Nicht zufällige Verteilung von Transkriptionsfaktor-Bindungsstellen im *Arabidopsis thaliana* Genom

Von der Fakultät für Lebenswissenschaften

der Technischen Universität Carolo-Wilhelmina

zu Braunschweig

zur Erlangung des Grades einer

Doktorin der Naturwissenschaften

(Dr. rer. nat.)

genehmigte

D i s s e r t a t i o n

von Claudia Galuschka

aus Wickede-Wimbern

1. Referent:	apl. Professor Dr. Reinhard Hehl
2. Referent:	Professor Dr. Dieter Jahn
eingereicht am:	16.06.2008
mündliche Prüfung (Disputation) am:	28.10.2008
Druckjahr	2008

Vorveröffentlichungen der Dissertation

Teilergebnisse aus dieser Arbeit wurden mit Genehmigung der Fakultät für Lebenswissenschaften, vertreten durch den Mentor der Arbeit, in folgenden Beiträgen vorab veröffentlicht:

Publikationen

Steffens N.O., Galuschka C., Schindler M., Bülow L., Hehl R. (2004) AthaMap: an online resource for *in silico* transcription factor binding sites in the *Arabidopsis thaliana* genome. Nucleic Acids Res. (32) D368-72.

Steffens N.O., Galuschka C., Schindler M., Bülow L., Hehl R. (2005) AthaMap web tools for database-assisted identification of combinatorial *cis*-regulatory elements and the display of highly conserved transcription factor binding sites in *Arabidopsis thaliana*. Nucleic Acids Res. (33) W397-402.

Bülow L., Steffens N.O., Galuschka C., Schindler M., Hehl R. (2006) AthaMap: from *in silico* data to real transcription factor binding sites. In Silico Biology (6) 0023.

Galuschka, C., M. Schindler, L. Bülow, R. Hehl. (2007) AthaMap web tools for the analysis and identification of co-regulated genes. Nucleic Acids Res. (35) D857-D862.

Tagungsbeiträge

Steffens, N. O., Galuschka, C., Bülow, L., Hehl, R. Towards a Transcription Factor Binding Site Map of *Arabidopsis thaliana*. Poster. 7th Int. Congress of Plant Molecular Biology, Barcelona, Spanien 2003

Steffens N.O., Galuschka C., Bülow L., Schindler M., Hehl R. AthaMap, an online resource for *in silico* transcription factor binding sites in the *Arabidopsis thaliana* genome. “German Conference on Bioinformatics“, München 2003, Deutschland

Galuschka C., Steffens N.O., Bülow L., Hehl R. Non-Random Distribution of Transcription Factor Binding Sites in the *Arabidopsis thaliana* Genome “ The Arabidopsis conference”, Berlin 2004, Deutschland

Steffens N.O., Galuschka C., Bülow L., Schindler M., Hehl R. AthaMap, an online resource for *in silico* transcription factor binding sites in the *Arabidopsis thaliana* genome. (Update) “The Arabidopsis conference”, Berlin 2004, Deutschland

Galuschka C., Steffens N.O., Bülow L., Hehl R. Non-Random Distribution of Transcription Factor Binding Sites in the *Arabidopsis thaliana* Genome (Update) “Botanikertagung 2004”, Braunschweig 2004, Deutschland

Steffens N.O., Galuschka C., Bülow L., Schindler M., Hehl R. AthaMap, an online resource for *in silico* transcription factor binding sites in the *Arabidopsis thaliana* genome. (Update) "Botanikertagung 2004", Braunschweig 2004, Deutschland

Hehl R., Bülow L., Cerff R., Galuschka C., Grohmann J., Hänsch R., Lisson R., Mendel R.R., Palme K., Sell S., Steffens N.O. Forward and reverse genetic approaches to study CO_2 dependent anaerobic signal transduction " 8th Conference of the International Society for Plant Anaerobiosis", Perth 2004, Australien

Steffens N.O., Galuschka C., Schindler M., Bülow L., and Hehl R. The AthaMap database resources for identification of *cis*-regulatory elements and the prediction of coregulated genes in *Arabidopsis thaliana*. "Plant Biology 2005", Seattle 2005, USA

Bülow L., Steffens N.O., Galuschka C., Schindler M. Hehl R., AthaMap: A genome-wide map of transcripton factor binding sites in *Arabidopsis thaliana*, Kiel 2006, Deutschland

Abkürzungsverzeichnis

bp	Basenpaare
A	Adenin
C	Cytosin
ChIP	Chromatin Immunoprecipitation
DNA	Desoxyribonucleic Acid
DTD	Document Type Definition
EMSA	Electrophoresis Mobility Shift Assay
G	Guanin
ID	Identifikationszeichen
IUPAC	International Union of Pure and Applied Chemistry
kb	Kilobasen
Mb	Megabasen
MS	Microsoft
NLS	Nuclear Localization S,ignal
RNA	Ribonucleic Acid
SQL	Structured Query Language
T	Thymin
TAIR	The Arabidopsis Information Resource
TIGR	The Institute for Genomic Research
TF	Transkriptionsfaktor
TFBS	Transkriptionsfaktor-Bindungsstelle
TSS	Transkriptionsstartpunkt
u.a.	unter anderem
UTR	nicht translatierte Region
XML	Extensible Markup Language
z.B.	zum Beispiel

Inhaltsverzeichnis

1. Einleitung

1.1 Transkriptionelle Regulation der Genexpression

In eukaryotischen Kernen liegt die genomische DNA mit Proteinen komplexiert vor. Dieser Komplex wird als Chromatin bezeichnet. Durch das Chromatin ist die DNA im Zellkern in gepacktem Zustand, was zu einer generellen Suppression der Genaktivität führt (Singh, 1998). Um die Genexpression zu ermöglichen, muss der Chromatin-DNA-Komplex in der Nähe des betreffenden Gens umgeformt werden, um den Zugang von Transkriptionsfaktoren, und um im Fall von proteinkodierenden Genen, das Rekrutieren des RNA-Polymerase II Transkriptionsinitiationskomplexes zu ermöglichen (Singh, 1998). Transkriptionsfaktoren sind DNA-bindende Proteine, die aktivierend oder auch inhibierend auf die Transkription wirken können. Einige Transkriptionsfaktoren sind in der Lage, den Chromatin-DNA-Komplex zu remodulieren (Singh, 1998). Dieser Komplex besteht aus DNA, die um Histone gewickelt ist, sowie aus Nichthiston-Proteinen. Der DNA-Histon-Komplex wird als Nucleosom bezeichnet und umfasst ca. 146 bp der DNA (Singh, 1998), wobei zwei Nucleosomen durch eine bis zu 60 bp lange Linker-DNA miteinander verbunden sind. Histone können von Enzymen modifiziert werden, wobei sie methyliert (Arginin-Reste), phosphoryliert (Serin-Reste) oder acetyliert (Lysin-Reste) werden können. Die Acetylierung sowie die Phosphorylierung führen in der Regel zu einer verminderten Bindungsfähigkeit der Histone an die DNA, was die Voraussetzung für die Transkription darstellt (Singh, 1998).

Regulatorische Transkriptionsfaktoren üben durch ihre Bindung an die DNA einen Einfluss auf die Chromatinstruktur aus und rekrutieren geeignete globale Koaktivatoren bzw. Korepressoren. Koaktivatoren enthalten beispielsweise Histon-Acetyl-Transferasen, mittels derer die Nukleosomenstruktur verändert werden kann, so dass eine Transkription möglich wird. Dabei entsteht ein transkriptionell aktiver Zustand, wodurch die Bindung weiterer Transkriptionsfaktoren ermöglicht wird. Umgekehrt wirken Korepressoren, die beispielweise Histon-Deacetyl-Transferasen enthalten, wodurch ein transkriptionell inaktiver Zustand entsteht (Singh, 1998).

Ein Gen ist ein Abschnitt auf der DNA, der bestimmte erblich bedingte Strukturen und Funktionen eines Organismus kodiert. Bei der Expression proteinkodierender Gene wird die DNA zunächst in mRNA transkribiert, wonach diese prozessiert und translatiert wird, so dass ein Protein gebildet wird.

Schematisch ist ein eukaryotisches Gen und sein Promotor in Abb. 1 dargestellt. In der Regel besteht ein Gen aus einer 5'UTR (Leader in Abb. 1), Exons, die durch Introns unterbrochen sein können, und einer 3'UTR (Trailer in Abb. 1), wobei die UTRs ebenfalls durch Introns unterbrochen sein können. Anders als in Abb. 1 werden in dieser Arbeit *cis*-regulatorische Sequenzen (Enhancer, Silencer) mit zum Promotorbereich gezählt.

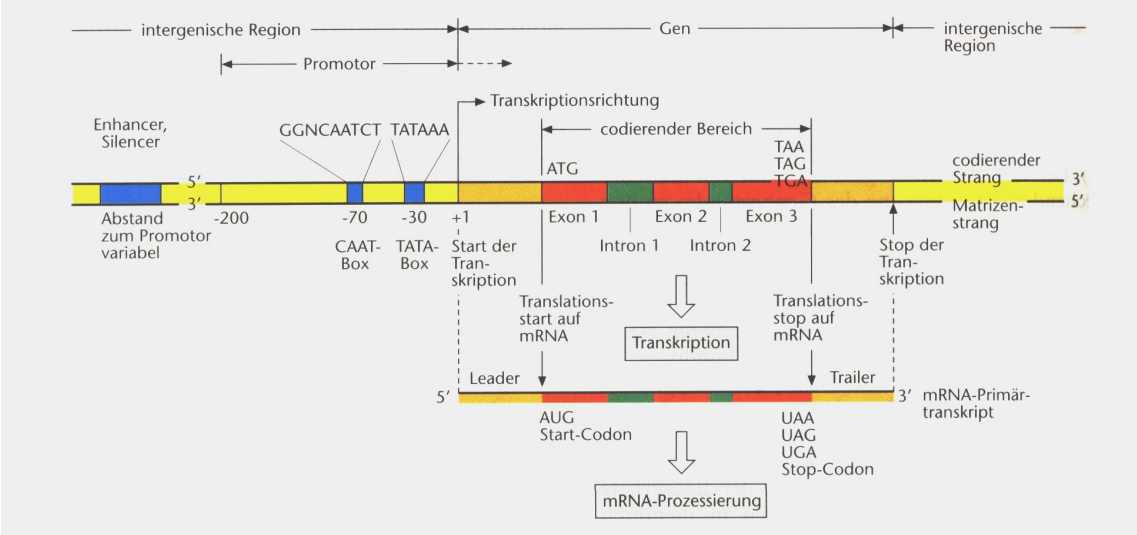


Abb. 1: Schematischer Aufbau eines eukaryontischen Gens und seines Promotors (Strasburger, 2002)

Viele Gene besitzen eine TATA-Box (Strasburger, 2002), an die zuerst der TFIID, welcher aus dem TATA-Box bindenden Protein (TBP) und einer Reihe von TAFs (TBP- associated factors) zusammengesetzt wird, bindet (Singh, 1998). Anschließend wird die RNA-Polymerase II rekrutiert, wonach schließlich das Gen transkribiert wird. Es gibt jedoch auch Promotoren ohne TATA-Box, die dann beispielsweise ein Initiatorelement (Inr, TCAGAGC) statt einer TATA-Box aufweisen, welches für Säuger belegt ist (Doan *et al.,* 2004).

Ziel bei der Aktivierung der Transkription von proteinkodierenden Genen ist es, die Bindung von basalen Transkriptionsfaktoren zu begünstigen und die RNA-Polymerase II zu rekrutieren, welche für die eigentliche Transkription eines Gens zuständig ist.
Die Genexpression wird durch die Bindung von Transkriptionsfaktoren an die DNA-Elemente, die vorwiegend im Promotor lokalisiert sind, reguliert (Reményi *et al.,* 2004). Im Gegensatz zu Prokaryonten, die oft einzelne Proteine für die transkriptionelle Regulation eines Gens aufweisen, beinhaltet die eukaryontische Regulation der Genexpression die Koordination multipler Proteine und ist daher kombinatorisch (Reményi *et al.,* 2004).

1.2 Transkriptionsfaktoren (TFen)

Transkriptionsfaktoren sind mehrere Domänen enthaltene Proteine, die an die DNA binden können und entweder aktivierend oder reprimierend auf die Transkription wirken (Liu *et al.,* 1999). Die meisten bekannten Transkriptionsfaktoren werden üblicherweise anhand ihrer DNA-Bindungsdomäne in Klassen eingeteilt (Jakoby *et al.,* 2002, Riechmann, 2000). In der Regel enthalten pflanzliche Transkriptionsfaktoren ein oder mehrere Kernlokalisierungssignale, eine die Transkription regulierende Domäne (Effektordomäne), eine Dimerisierungsdomäne sowie mindestens eine DNA-Bindungsdomäne (Liu *et al.,* 1999) (Abb. 2). Einigen Transkriptionsfaktoren fehlt die Effektordomäne oder aber die DNA-Bindungsdomäne (Liu *et al.,* 1999). Auf die einzelnen Charakteristika wird im Folgenden näher eingegangen.

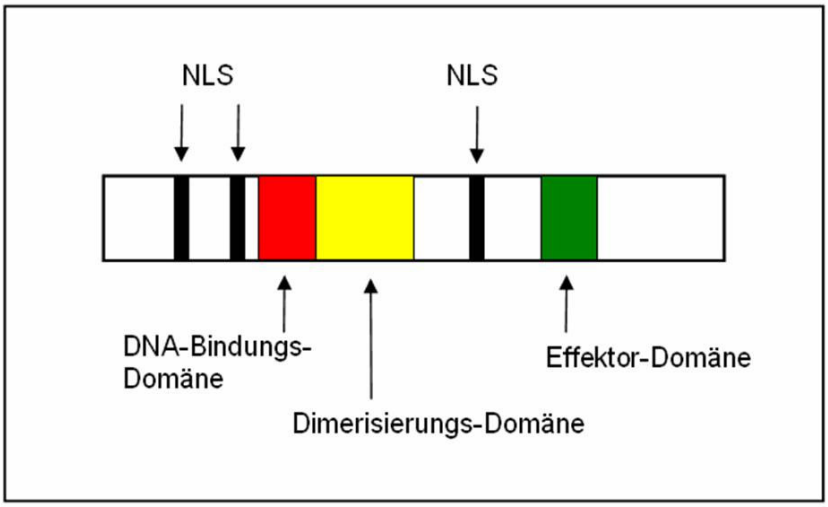


Abb. 2: Schematische Darstellung der funktionalen Domänen eines pflanzlichen Transkriptionsfaktors (nach Liu *et al.,* 1999). NLS = Kernlokalisierungssignal, welches auch innerhalb der DNA-Bindungs-Domäne oder der Dimerisierungsdomäne lokalisiert sein kann, weitere Erläuterungen siehe Text.

Das Kernlokalisierungssignal bewirkt, dass die Transkriptionsfaktoren vom Cytoplasma, dem Ort der Translation, in den Zellkern transportiert werden können. Hierzu muss das Protein durch die Poren des Zellkerns gelangen. Die nuklearen Poren von höheren Pflanzen enthalten Proteine, die an das Kernlokalisierungssignal von Transkriptionsfaktoren binden kann. Das Kernpeptid ist mit Arginin (R)- und Lysinresten (K) angereichert, wobei die Anzahl dieser Kernlokalisierungssignale in der Sequenz und in der Anzahl variieren kann und auch das Auftreten von nur einem Kernlokalisierungssignal vorkommen kann (Liu *et al.,* 1999). Es ist wahrscheinlich, dass einige pflanzliche Transkriptionsfaktoren auch ohne ein

Kernlokalisierungssignal durch Dimerisierung mit anderen Transkriptionsfaktoren, die dieses Signal aufweisen, in den Kern transportiert werden können (Liu *et al.,* 1999).

Ein weiteres allgemeines Charakteristikum eines Transkriptionsfaktors ist die DNA-Bindungsdomäne. Eine DNA-Bindungsdomäne erkennt kurze, etwa 5 bis 15 Nukleotid lange DNA-Sequenzen (*cis*-agierende Elemente) und ist für die spezifische Bindung des Transkriptionsfaktors an die DNA verantwortlich (Reményi *et al.,* 2004). Diese Domäne enthält oft basische Reste, die in vielen Fällen hoch konserviert sind. Dies ist sinnvoll, da die DNA negativ geladen ist. In der Regel weisen pflanzliche Transkriptionsfaktoren eine oder mehrere DNA-Bindungsdomänen auf. Hierbei kann ein Transkriptionsfaktor auch über eine spezifische und eine nichtspezifische DNA-Bindungsdomäne verfügen, wie dies für VP1 (VIVIPAROUS 1) demonstriert werden konnte (Liu *et al.,* 1999). Dieser Faktor besitzt die sequenzspezifische DNA-Bindungsdomäne BR3 sowie die nicht sequenzspezifische DNA-Bindungsdomäne BR2. Es konnte gezeigt werden, dass VP1 das Em-Gen aus dem Weizen sogar über die nicht spezifische DNA-Bindungsdomäne aktiviert. Diese BR3 lose Regulation benötigt Mitglieder der Proteinfamilie 14-3-3, denen jedoch die DNA-Bindungsdomäne fehlt. Jedoch vermitteln diese eine Bindung zwischen VP1 und einem *cis*-Element-spezifischen Transkriptionsfaktor, zum Beispiel EmBP1, an die Promotorregion (Liu *et al.,* 1999).

Zusätzlich können Transkriptionsfaktoren auch Domänen für die Interaktion mit anderen Proteinen aufweisen, wobei Homo- und/oder Heterodimere gebildet werden können, was einen Einfluss auf die DNA-Bindungsspezifität und auf die Affinität des Transkriptionsfaktors zur DNA hat (Jakoby *et al.,* 2002). Mitglieder der nichtsteroiden Kernrezeptor Superfamilie sind in der Lage, an HREs (hormone response elements) zu binden. HREs bestehen aus den zwei Hexanucleotiden AGGTCA, wobei die Orientierung und der Abstand der beiden Motive die Identität und die Form der Bindung (Monomer, Homodimer, Heterodimer) der Proteine bestimmt. Die Orientierung der DNA-Sequenzen kann gleich (→ →), aber auch gegensätzlich (← → bzw. → ←) sein (Reményi *et al.,* 2004). Bei anderen Transkriptionsfakoren, wie bei Mitgliedern der Klasse bZIP, ist die Bildung von Homo- bzw. Heterodimeren für die Fähigkeit notwendig, an die DNA zu binden (Jakoby *et al*., 2002, Singh, 1998). Auch die Kernlokalisierung kann von der Art der Dimerisierung beeinflusst werden (Liu *et al.,* 1999). Durch eine Interaktion von Transkriptionsfaktoren können regulatorische DNA-Sequenzen sogar in eine räumliche Nähe zueinander gebracht werden, wie das folgende Beispiel zeigt. Für das Gen cyllla aus dem Seeigel konnte demonstriert werden, dass regulatorische Sequenzen, die in einiger Entfernung des basalen Transkriptionsapparates liegen, DNA-Bindungssequenzen enthalten, an die Transkriptionsfaktoren binden, welche die DNA durch Multimerisierung nach der Bindung

durch einen Loop in räumliche Nähe zu anderen regulatorischen Sequenzen bringen, und damit auch näher an den basalen Transkriptionsapparat (Istrail und Davidson, 2004), wodurch die Transkription aktiviert wird.

Ein weiteres Charakteristikum von Transkriptionsfaktoren ist das Vorhandensein von Effektordomänen. Effektordomänen vermitteln die Aktivierung oder Reprimierung von Zielgenen (Blancafort *et al.,* 2004), wobei dasselbe Protein auch beide Funktionen aufweisen kann (Singh, 1998). Die Wirkung eines Transkriptionsfaktors auf die Genexpression ist unabhängig von seiner Klassenzugehörigkeit.

Eine Repression könnte durch kompetitive Bindung von Transkriptionsfaktoren an dieselben DNA-Elemente, durch eine Maskierung der regulatorischen Domäne durch Dimerisierung von Transkriptionsfaktoren oder durch eine Interaktion der reprimierenden Domäne mit anderen Transkriptionsfaktoren auftreten (Liu *et al.,* 1999). Relativ wenig ist über die Struktur und die Funktionsweise von Aktivierungsdomänen bekannt. Einige können anhand ihres vermehrten Auftretens bestimmter Aminosäurereste klassifiziert werden. Es gibt z.B. saure, Gln-, Pro- und Ser/Thr- reiche Aktivierungsdomänen (Singh, 1998).

Die Funktionalität der Transkriptionsfaktoren wird vorwiegend post-translational beeinflusst. Prinzipiell gelangen diese Proteine aufgrund ihres Kernlokalisierungssignals in den Zellkern und haben mittels ihrer Domänen eine Wirkung auf die Transkription. Transkriptionsfaktoren können jedoch modifiziert werden. Das Kernlokalisierungssignal kann beispielsweise durch Phosphorylierung direkt modifiziert oder auch durch intra- und intermolekulare Maskierung beeinflusst werden, so dass das Protein gar nicht erst in den Zellkern gelangen kann (Liu *et al.,* 1999). Die Phosporylierung bzw. Dephosporylierung von Proteinen kann einen Einfluss auf die Fähigkeit des Transkriptionsfaktors ausüben, an die DNA zu binden, wodurch die Genexpression ebenfalls beeinflusst werden kann, wobei eine Phosphorylierung einen positiven Einfluss auf die Bindungsfähigkeit haben kann (Liu *et al.,* 1999). Dies lässt darauf schließen, dass die Funktion eines Transkriptionsfaktors von vielen weiteren Gegebenheiten in der Zelle abhängt.

Nachdem auf die einzelnen Charakteristika eines typischen Transkriptionsfaktors eingegangen wurde, folgt nun eine Übersicht über die verschiedenen Klassen von pflanzlichen Transkriptionsfaktoren. Hierbei liegt der Fokus auf denjenigen Klassen, die für die diese Arbeit relevant sind. Zusätzlich zu den beschriebenen Klassen wurde ein Faktor der Klasse C2H2(Zn) verwendet.

1.2.1 Klassen von pflanzlichen Transkriptionsfaktoren

Transkriptionsfaktoren weisen verschiedene Domänen auf und werden üblicherweise anhand ihrer DNA-Bindungsdomäne in Klassen oder auch Familien eingeteilt. DNA-Bindungsdomänen derselben Klasse erkennen oftmals dieselbe DNA-Kernsequenz, die konserviert sein muss, damit der Faktor an die DNA binden kann. Die von dem Faktor gebundene DNA-Sequenz ist länger als die Kernsequenz, jedoch können die flankierenden Nukleotide variieren.

Transkriptionsfaktoren der Klasse bZIP besitzen eine basische DNA-Bindungsdomäne, wobei die Proteine vorwiegend die DNA-Kernsequenz ACGT erkennen und an sie binden (Jakoby *et al.,* 2002). Die Faktoren haben noch eine weitere ihnen gemeinsame Domäne, die Dimerisierungsdomäne, welche im Fall der bZIP Faktoren der Leucin-Zipper ist. bZIP Transkriptionsfaktoren bilden Homodimere oder Heterodimere, wobei die Dimerisierung die Voraussetzung für die DNA-Bindung ist. Das Vorkommen von bZIP Faktoren ist nicht auf Pflanzen beschränkt. Sie sind z.B. auch in *Drosophila melanogaster*, *Caenorhabditis elegans* und *Saccharomyces cerevisiae* vertreten (Riechmann *et al.,* 2000). Die Funktionen von bZIP Faktoren in *Arabidopsis thaliana* umfassen eine Mitwirkung bei Stressantworten oder Pathogenabwehr, aber sie spielen ebenfalls bei der Blütenentwicklung und Samenreifung eine Rolle (Jakoby *et al.,* 2002).

bHLH (Basic Helix-Loop-Helix) Faktoren besitzen wie die bZIP Transkriptionsfaktoren ebenfalls eine basische DNA-Bindungsdomäne, die wie bZIP Faktoren die DNA-Kernsequenz ACGT erkennt. Sie können ebenfalls Homo- und Heterodimere bilden. bHLH Transkriptionsfaktoren sind beispielsweise in den Lichtsignalweg involviert (Duek und Fankhause, 2005), wobei sie als negative und seltener als positive Regulatoren fungieren. Das Vorhandensein von bHLH Transkripten steht unter der Zellzyklus-Kontrolle. Transkriptionsfaktoren der Klasse bHLH kommen nicht nur bei Pflanzen vor, sie sind beispielsweise auch in *Drosophila melanogaster*, *Caenorhabditis elegans* und *Saccharomyces cerevisiae* vertreten (Riechmann *et al.,* 2000).

WRKY Transkriptionsfaktoren wurden bislang in Pflanzen und zwei Vertretern der Schleimpilze gefunden. Sie erkennen meist die sogenannte W-Box, welche eine DNA-Sequenz (C/T)TGAC(T/C) aufweist. Jedoch kann die vom Faktor gebundene Sequenz auch variieren (Ulker und Somssich, 2004). WRKY Faktoren sind z.B. in die Pathogenabwehr, Seneszenz, Verwundungsreaktionen und Reaktionen auf abiotischen Stress involviert und können aktivierend oder auch reprimierend wirken (Eulgem *et al.,* 2000).

Transkriptionsfaktoren der Klasse DOF (DNA-Binding With One Finger, C2C2(Zn)) sind in Stressantworten involviert, in den Phytochrom- Signalweg, in Lichtantworten und in Reaktionswege bei Anwesenheit pflanzlicher Hormone. An der Samenentwicklung sind sie ebenfalls beteiligt (Yanagisawa, 2004). DOF Faktoren sind pflanzenspezifisch und weisen eine konservierte DNA-Bindungsdomäne auf, die die konservierte DNA-Kernsequenz (A/T)AAAG erkennt (Yanagisawa, 2002), wobei nicht alle AAAG Promotorsequenzen tatsächliche Zielsequenzen für DOF Faktoren darstellen (Yanagisawa, 2004). Transkriptionsfaktoren der Klasse DOF können aktivierend oder reprimierend auf die Transkription wirken (Yanagisawa, 2004). Bekannt ist, dass die DOF-Domäne selbst die Interaktion mit anderen Proteinen ermöglicht, wobei diese für Mitglieder der Klasse bZIP und auch HMG-Proteine (High Mobility Group) gezeigt werden konnte (Yanagisawa, 2004).

Trihelix Transkriptionsfaktoren sind pflanzenspezifisch und sind z.B. an der Regulation von Genen, die für die Lichtantwort zuständig sind, beteiligt. Des Weiteren können sie eine Rolle bei der Morphogenese und der TIA (Terpenoid Indole Alkaloid) Biosynthese einnehmen (Brewer *et al.,* 2004, Ouwerkerk *et al.,* 1999). Sie erkennen die DNA-Sequenz GT(A/T)AA(A/T)(A/T) (Ouwerkerk *et al.,* 1999).

Transkriptionsfaktoren der Klasse TCP sind pflanzenspezifisch, sie sind an der Morphogenese von Pflanzen beteiligt (Navaud *et al.,* 2007). Sie können je nach Subklasse eher aktivierend oder reprimierend auf die Transkription wirken (Ingram und Waites, 2006). Die Erkennungssequenz variiert je nach Subklasse und umfasst für die Klasse I die DNA-Sequenz GGNCCCAC und GTGGNCCC für die Klasse II (Kosugi und Ohashi, 2002). Bekannt ist, dass TCP Proteine mit NAC Transkriptionsfaktoren interagieren können (Ingram und Waites, 2006).

Die Mitglieder von MADS Transkriptionsfaktoren kommen nicht nur in Pflanzen sondern auch in Hefen und Säugern vor und sind in pflanzlichen Systemen wichtige Regulatoren bei der Blütenentwicklung (Jack, 2001). Sie können Homo- und Heterodimere bilden und können zusätzlich einen ternären Komplex mit Dimeren der Klasse MADS bilden, wobei die Dimere direkt miteinander interagieren und die zwischen den beiden Bindungsorten liegende DNA in einer Schlaufe vorliegt (Ferrario *et al.,* 2003). Viele MADS Gene regulieren ihre eigene Expression (Rijpkema *et al.,* 2007). Transkriptionsfaktoren der Klasse MADS erkennen vorwiegend die DNA-Sequenz $CC(A/T)_6GG$, die als CArG Motiv bezeichnet wird (Tang und Perry, 2003).

AP2/EREBP Transkriptionsfaktoren kommen nur in Pflanzen vor (Riechmann *et al.,* 2000). Die Einteilung in die beiden Subklassen AP1 und EREBP erfolgt mittels der Anzahl der vorhandenen AP2-DNA-Bindungsdomänen der Faktoren. AP2 Faktoren weisen zwei AP2 Domänen auf, während EREBP Transkriptionsfaktoren nur eine AP2 Domäne beinhalten. Transkriptionsfaktoren der Klasse AP2 sind z.B. in Prozesse der Blütenentwicklung und bei der Samenentwicklung involviert, bei EREBP Faktoren sind diese z.B. an der Ausbildung von Krankheitsresistenzen, wie der spezifischen Pathogenabwehr, beteiligt (Gutterson und Reuber, 2004, Riechmann und Meyerowitz, 1998). Diese AP2/EREBP Klasse weist eine weitere Subklasse auf, die RAV ähnlichen Transkriptionsfaktor, deren Mitglieder zusätzlich zu der AP2 Domäne eine B3 Domäne aufweisen (Gutterson und Reuber, 2004). Die Zielsequenz von EREBP Proteinen ist die CGG box (AGCCGCC), bei AP2 Faktoren ist eine gemeinsame DNA-Sequenz bislang nicht einheitlich beschrieben worden. Dies gilt auch für RAV Faktoren. Im Falle von RAV1 beispielsweise sind zwei separate DNA-Sequenzen beschrieben worden, an die der Transkriptionsfaktor bindet, CAACA sowie CACCTG (Kagaya *et al.,* 1999).

Faktoren der Klasse MYB kommen nicht nur in Pflanzen sondern z.B. auch in *Drosophila melanogaster*, *Caenorhabditis elegans* und *Saccharomyces cerevisiae* vor (Riechmann *et al.,* 2000). Sie können entweder zwei oder drei nicht perfekte Wiederholungen der MYB-Domäne aufweisen und sind z.B. an der Kontrolle des epidermalen Zellschicksals beteiligt (Ramsay und Glover, 2005). Die häufigste DNA-Kernsequenz, an die die Faktoren binden, lautet AAC. Zwei bestimmte Sequenzen, an die einige MYB Faktoren mit zwei MYB-Domänen binden können, wurden als MYB Bindingssite I, (T/C)AAC(G/T)G(A/C/T)(A/C/T), und MYB Bindingsite II (TAACTAAC) benannt (Jin und Martin, 1999).

Das Vorkommen von GATA Transkriptionsfaktoren ist nicht auf Pflanzen beschränkt, diese kommen auch in anderen Eukaryonten vor (Reyes *et al.,* 2004). Sie erkennen in der Regel die Kernsequenz (A/T)GATAR, und im Falle von *Arabidopsis* kann auch NGATNN als Kernsequenz gebunden werden. In Pflanzen sind Mitglieder dieser Klasse z.B. am Lichtübertragungsweg beteiligt (Naito *et al.,* 2007), sie spielen aber auch eine Rolle bei der Kontrolle der circadiaren Rhythmik und dem Stickstoff Metabolismus (Reyes *et al.,* 2004).

HD (Homeodomain) Transkriptionsfaktoren sind laut Riechmann *et al.*, 2000, nicht pflanzenspezifisch. Sie wurden auch in Tieren und Pilzen identifiziert und sind für die Festlegung des Zellschicksals notwendig (Reiser *et al.,* 2000). Sie erkennen vorwiegend die DNA-Sequenz ATTA (Chan *et al.,* 1999). Den Faktoren gemeinsam ist die Homeo-Domäne, jedoch weisen sie je nach Subklasse weitere Domänen auf. HD-Knotted

Transkriptionsfaktoren besitzen zusätzlich zur HD Domäne die ELK sowie eine MEINOX Domäne. Beide ermöglichen wahrscheinlich die Interaktion mit anderen Proteinen (Reiser *et al.,* 2000). Exprimiert werden diese Gene z.B. in Meristemen, nicht aber in lateralen Organen (Reiser *et al.,* 2000). Eine den HD-Knotted gemeinsame Ziel-DNA-Sequenz wurde bislang nicht beschrieben. Im Falle von HVH21 (Knox3) z.B. umfasst sie TGAC (Krusell *et al.,* 1997).
Eine weitere Subklasse der HD Transkriptionsfaktoren stellen HD-PHD Proteine dar, die außer der HD Domäne eine PHD (Plant Homeodomain) Domäne aufweisen, wobei letztere als Protein-Protein-Interaktionsdomäne postuliert wird (Chan *et al.,* 1999). Die Proteine spielen z.B. bei Fortpflanzungsprozessen eine Rolle (Chan *et al.,* 1999) und wirken bei der Chromatin-vermittelten Transkriptionskontrolle mit (Mellor, 2006).
HD-Zip Transkriptionsfaktoren weisen neben der HD-Domäne eine ZIP Domäne auf, die bekanntermaßen für die Protein-Protein-Interaktion zuständig ist (Chan *et al.,* 1999). Die Faktoren binden vorwiegend an die DNA-Sequenz CAAT(A/T)ATTG, wobei sie, als Voraussetzung für die Bindung an die DNA, Dimere bilden (Chan *et al.,* 1999). Beteiligt sind die Faktoren bei der Regulation der Blattentwicklung, sowie beim Signaltransduktionsweg als Folge eines Wasserdefizits (Chan *et al.,* 1999). Einige dieser Faktoren konnten in pflanzlichen Embryonen nachgewiesen werden (Byrne, 2006). Bekannt ist, dass einige einen Effekt auf die Festlegung der Xylem-Phloem-Identität ausüben (Demura und Fukuda, 2006).

1.2.2 Bindungsstellen von Transkriptionsfaktoren im Promotor und innerhalb von Genen

Transkriptionsfaktoren können durch ihre Bindung an DNA-Sequenzen aktivierend oder reprimierend auf die Transkription wirken. Die Bindung von Transkriptionsfaktoren ist vorwiegend im Promotorbereich von Genen lokalisiert (Reményi *et al.,* 2004).
Als Beispiel für Bindungsstellen in einem Promotor sei hier das 22kD-Zein Gen aus Mais erwähnt. Zein gehört zu den Speicherproteinen, die während der Samenproduktion gebildet werden, und dient dem Embryo als Stickstoffquelle. Die Expression findet im Endospermgewebe, welches ein Nährgewebe für den sich entwickelten Embryo ist, statt. Im Promotorbereich des Gens befinden sich eine O2 sowie eine P Box, wobei die O2 Box näher am Transkriptionsstart liegt als die P Box, und beide etwa 20 Nukleotide entfernt voneinander lokalisiert sind. An die O2 Box bindet das O2 Dimer der Klasse bZIP, an die P Box der Transkriptionsfaktor PBF der Klasse DOF. Es konnte gezeigt werden, dass O2 aktivierend wirkt und dass dazu eine funktionale P Box notwendig ist. Daher ist die Bindung beider Faktoren für die Aktivierung der Transkription des Gens notwendig (Singh, 1998).

Es gibt jedoch auch Beispiele für die Lokalisierung von Transkriptionsfaktor-Bindungsstellen, die nicht im Promotorbereichen, sondern innerhalb von transkribierten Bereichen vorkommen und regulierend auf die Transkription wirken. Beispielsweise können Bindungsstellen in der 5'UTR vorliegen. Beim CPRF1 (Common Plant Regulatory Factor 1) Transkriptionsfaktor der Klasse bZIP aus der Petersilie befinden sich in der 5'UTR des Gens zwei benachbarte ACGT Motive, deren palindromische Zentren eine Distanz von 12 Nukleotiden aufweisen. *In vivo* Analysen zeigten, dass dieser Faktor durch die Interaktion mit diesen Elementen wahrscheinlich negativ autoregulierend wirkt (Feldbrügge *et al.,* 1994). Wird dieses Element vor einen Minimalpromotor und ein Reportergen kloniert, kommt es zu einer Verstärkung der Transkription.

Für die Lokalisierung von Transkriptionsfaktor-Bindungsstellen in der 5'UTR gibt es auch Beispiele bei nichtpflanzlichen Organismen. *In vitro* Studien sowie Reportergenanalysen konnten belegen, dass der RLD-1 Kernrezeptor an ein „fettspezifisches" Element im FAS (Fatty Acid Sythase) Gen aus der Ratte in Gegenwart des Retinoidrezeptors RXR, ebenfalls einem Protein aus der Ratte, welches mit dem Faktor ein Heterodimer bildet, bindet und die Genexpression aktiviert. Die DNA-Sequenz ist in der 5'UTR, 271 bp upstream des Translationsstartpunktes, lokalisiert und weist die beiden Sequenzen (G/A)GGTCA, getrennt durch vier Nukleotide, auf (Apfel *et al.,* 1994). Ein weiteres Beispiel für Bindungsstellen in der 5'UTR liefert das Haushaltsgen Dihydrofolate Reduktase (DHFR) aus der Maus. Hier konnte durch *in vivo* Analysen demonstriert werden, dass eine Deletion von 46 Nukleotiden in der 5'UTR eine starke Reduzierung der Transkriptionsaktivität zur Folge hatte, woraus die Schlussfolgerung resultiert, dass dieser regulatorische Bereich für die Aktivierung der Transkription des Gens notwendig ist. Der an diese Sequenz bindende Transkriptionsfaktor wurde bei diesen Analysen nicht identifiziert (Farnham und Means, 1990).

Regulatorische DNA-Sequenzen können auch in Introns liegen, wobei die die Genexpression regulierenden Introns gewöhnlich in der Nähe der 5' Region von Genen lokalisiert sind und einen längeren Nukleotidbereich umfassen (Jeong *et al.,* 2006).
Beim MADS-Box Gen AG (AGAMOUS) aus *Arabidopsis*, welches z.B. in die Blütenentwicklung involviert ist, ist bekannt, dass die negative Regulation des Transkriptionsfaktors durch Bindung der Transkriptionsfaktoren AP2 (APETALA2) bzw. LUG (LEUNIG) an *cis*-Elemente im zweiten Intron, welches etwa 3 kb groß ist, erfolgt. Das zweite Intron befindet sich downstream des ersten Exons, da in der 5'UTR ebenfalls ein Intron vorhanden ist. AP2, ein Faktor der Klasse AP2/EREBP, und der Transkriptionsfaktor LUG, welcher eine WD40 Domäne aufweist, wirken dabei teilweise in unterschiedlichen Geweben (Sieburth und Meyerowitz, 1997). Der Transkriptonsfaktor LFY hingegen aktiviert die AG

Transkription durch Bindung an Enhancer-Sequenzen im zweiten Intron (Busch *et al.,* 1999). Es konnte also demonstriert werden, dass verschiedene DNA-Sequenzen in ein und demselben Intron entweder zu einer Aktivierung oder Reprimierung der Genexpression führen können.

Im Fall von *Arabidopsis* sind noch weitere Beispiele für regulatorische Sequenzen in Introns bekannt. Für AGL15 (AGAMOUS-like 15) aus *Arabidopsis* konnte mittels ChIP (Chromatin Immunoprecipitation) Analysen gezeigt werden, dass der Transkriptionsfaktor an eine 16 bp lange DNA-Sequenz im einzigen Intron von DTA4 (Downstream Target of AGL15) binden kann. Dies führt *in vivo* zu einer erhöhten gewebespezifischen Genexpression von DTA4 (Tang und Perry, 2003).

Reportergenanalysen mit den Promotoren der *Arabidopsis thaliana* Gene PRF1 und PRF2, zwei kleinen Actin-bindenden Profilinen, die in die Regulation des Actin Zytoskeletts involviert sind, konnten belegen, dass die Anwesenheit des ersten Introns der beiden Gene die Genexpression verstärkt (Jeong *et al.,* 2006). Das Vorhandensein regulatorischer Introns wird auch bei anderen Pflanzenarten, wie z.B. Tabak, vorgefunden. Bindet der Transkriptionsfaktor NtH15 (*Nicotiana tabacum* Homeobox 15) an eine DNA-Sequenz im ersten Intron des Gens Ntc12, einer GA 20-Oxidase, hat dies eine Reprimierung der Genexpression zur Folge. Das Genprodukt von Ntc12 oxidiert Gibberellin, ein pflanzliches Hormon, das beispielsweise das Stammwachstum beschleunigt. Durch die Oxidierung wird die Substanz in eine physiologisch inaktive Form überführt. Die Reprimierung der Transkription von Ntc12 hat zur Folge, dass es zu einem Ansteigen des Gibberellin-Spiegels kommt (Sakamoto et al., 2001).

Die Beispiele für regulatorische Introns sind nicht nur auf Pflanzen beschränkt. Unter Verwendung von ChIP sowie funktionaler Analysen konnte beim humanen Gen SCN3B (Sodium Channel Subunit Beta 3) 10 kb upstream vom ersten Exon sowie auch im dritten Intron je eine DNA-Bindungsstelle für den Transkriptionsfaktor p53, der als Tumorsupressor eine wichtige Verwendung bei der Krebstherapie findet, identifiziert werden, wobei p53 bei SCN3B als positiver Regulator fungiert (Adachi *et al.,* 2004).

Einen weiteren Beleg für eine negativ regulierende Funktion von Introns liefern die folgenden beiden Beispiele. Das Gen GFI1 (Growth Factor Independence-1), ein Transkriptionsfaktor aus der Maus, wird zu den Onkogenen gezählt und spielt bei der Blutbildung eine Rolle. Im ersten Intron dieses Genes liegen regulatorische Bereiche, an die der Transkriptionsfaktor selbst binden kann und seine eigene Expression inhibiert. Für die funktionalen Analysen wurde zwar die genomische Sequenz des homologen GFI1 Gens aus der Ratte eingesetzt, jedoch ist die Bindungssequenz im ersten Intron in der Ratte (aber nicht beim Menschen)

konserviert (Doan *et al.,* 2004). Das zweite Beispiel ist das erste Intron des Gens GRIK5 (Glutamate Receptor Ionotropic Kainate 5) aus der Ratte, das Bindungssequenzen aufweist, die eine Repression der Genexpression *in vivo* bewirken (Chew *et al.,* 1999).

Ein weiteres interessantes Beispiel für das Vorhandensein regulatorischer proximal zum Transkriptionsstart gelegener Introns, wobei hier mehr als eine Transkriptionsfaktor-Bindungsstelle in diesem Bereich aufzufinden ist, liefert das humane Apolipoprotein B Gen, welches insgesamt 29 Exons und 28 Introns enthält. B Apolipoproteine sind Plasmaproteine, die eine entscheidende Rolle im Säuger-Lipoprotein-Metabolismus spielen. Durch Footprint Analysen konnten die exakten DNA-Sequenzen, an die Transkriptionsfaktoren gebunden vorlagen, identifiziert werden. Insgesamt wurden 5 Bindungssequenzen im ersten Intron und 6 Bindungssequenzen im zweiten Intron ermittelt. *In vivo* Experimente konnten belegen, dass die Sequenzen im ersten und zweiten Intron gewebespezifisch als Enhancerelemente wirken (Brooks *et al.,* 1991).

Regulatorische Bereiche können auch am Ende von Genen in der 3'UTR auftreten, wie die folgenden Beispiele belegen. Für die beiden Nichthiston Proteine HMG-I und HMG-Y aus dem Menschen, die wie Transkriptionsfaktoren aufgebaut sind und die an die DNA binden können, konnte *in vitro* gezeigt werden, dass beide innerhalb einer 300 bp langen DNA-Sequenz der 3'UTR des Rinder Interleukin-2 Gens (BLT) binden können (Banks *et al.,* 2000). Das Genprodukt von VEGF (Vascular Endothelial Growth Factor), ein wirksamer Stimulator bei der Gefäßbildung, ist ein Indikator für viele Tumore, z.B. in der Brust und im Uterus. Die Transkription des Gens wird über den Östrogenrezeptor vermittelt. *In vivo* Studien konnten belegen, dass in der 3'UTR die Sequenz GAGCAcccTGCCC aus dem VEGF Gen der Ratte in Gegenwart des menschlichen Östrogenrezeptors (ER-α sowie ER-β) eine Enhancerfunktion vermittelt. Eine weitere Bindungssequenz für den Östrogenrezeptor ist in der 5'UTR lokalisiert (GGGCAaagTGACT), wobei das Induktionsmuster dieses Elements komplexer ist (Hyder *et al.,* 2000).

Somit liegen zahlreiche Beispiele für Transkriptionsfaktor-Bindungsstellen in nicht-kodierenden Bereichen, also im Promotorbereich, in der 5'UTR, in Introns und der 3'UTR vor. Funktionelle Bindungsstellen in Exons, die für Proteine kodieren und somit einem stärkeren evolutionären Druck unterliegen, konnten bislang nur in einem Fall (Merriam und Chess, 2007) identifiziert werden.

1.2.3 Identifizierung von Transkriptionsfaktor-Bindungsstellen (TFBSn)

Nachdem auf Transkriptionsfaktoren und deren Bindungsorte an die DNA eingegangen wurden, bleibt die Frage offen, wie Transkriptionsfaktor-Bindungsstellen als solche experimentell nachgewiesen werden können. Beispielhaft werden einige Methoden aufgeführt.

Eine Möglichkeit, eine DNA-Bindung von Proteinen nachzuweisen, ist die Verwendung eines Zellextrakts, welches mit synthetisierten und markierten Oligonukleotiden vereinigt und auf einem Gel aufgetragen wird. Jedoch ist dann nicht klar, welcher Transkriptionsfaktor an die DNA binden kann. Um einen Transkriptionsfaktor identifizieren zu können, kann dieser z.B. in *E. coli* exprimiert und mit diesen Oligonukleotiden zusammengeführt werden. Hierbei ist es für die Qualität der Aussage, dass der Faktor an die zur Verfügung gestellte DNA binden kann, auch entscheidend, die spezifische Bindung an diese DNA nachweisen zu können, daher werden Kompetitionsexperimente angeschlossen. Bei diesen wird zum einen zusammen mit den markierten Oligonukleotiden eine z.B. zufällige DNA-Sequenz im Überschuss hinzugegeben. Somit kann der Nachweis erbracht werden, dass die Bindung des Transkriptionsfaktors spezifisch ist. Andererseits wird neben dem markierten Oligonukleotid ein Überschuss an dem selben, aber diesmal nicht markierten Oligonukleotid hinzugefügt, wodurch sich dann die Stärke der beobachteten Bande im Gel deutlich verringert, wenn der Transkriptionsfaktor fähig ist, an die DNA-Sequenz zu binden. Zusammengefasst wird diese Methode des Nachweises verschiedener Transkriptionsfaktor-Bindungsstellen als EMSA (Electrophoresis Mobility Shift Assay) bezeichnet (Chinnusamy *et al.,* 2003). Um zu einer Sammlung von Transkriptionsfaktor-Bindungsstellen zu gelangen, können eine Vielzahl von verschiedenen synthetisierten und markierten Oligonukleotiden eingesetzt werden. Ist die Transkriptionsfaktor-Klasse bekannt, können DNA-Sequenzen, die dieselbe Kernsequenz aufweisen, sich aber in den flankierenden Nukleotiden unterscheiden, synthetisiert werden, um die Anzahl der einzusetzenden Oligonukleotide einzuschränken.

Aus diesen Experimenten geht jedoch nicht hervor, welche Nukleotide vom Transkriptionsfaktor direkt gebunden werden. Um dies feststellen zu können, kann z.B. ein sogenannter „DNAse footprint“ durchgeführt werden. Das bedeutet, dass neben der DNA-Sequenz, an die der Transkriptionsfaktor binden kann, und dem Transkriptionsfaktor selbst, DNAse hinzugefügt wird, die die nicht durch die Bindung des Transkriptionsfaktors geschützte DNA degradiert. Anschließend erfolgt eine Sequenzier-Gelelektrophorese. Die genaue Lokalisierung der Bindungssequenz kann ermittelt werden, wenn parallel zu dem beschriebenen Ansatz unbehandelte DNA im Gel aufgetragen wird (Duval *et al.,* 2002).

Bislang wurde allein der Nachweis der Bindung eines Transkriptionsfaktors an die DNA behandelt (*in vitro* Experimente). Um zu untersuchen, ob die Bindungsstellen in der Zelle (*in vivo*) auch tatsächlich von dem Transkriptionsfaktor erkannt werden, und es zu einer Auswirkung auf die Transkription kommt, können funktionale Analysen mit Reportergenen eingesetzt werden. Hierzu kann z.B. das „Yeast One Hybrid" System verwendet werden. Einerseits wird ein Fusionsprotein des Transkriptionsfaktors mit der Aktivierungsdomäne von z.B. GAL4 erzeugt. Andererseits wird die zu testende DNA-Sequenz vor einen Minimalpromotor und ein Reportergen, hier LacZ, ligiert. Die Konstrukte werden in Hefe transformiert. Kann der Transkriptionsfaktor durch Bindung an die DNA die Transkription aktivieren, kann eine *β*-Galactosidase Aktivität festgestellt werden, die auf eine Aktivierung der Transkription durch den eingesetzten Transkriptionsfaktor hindeutet (Teakle *et al.,* 2002).

Die Bindung von Transkriptionsfaktoren an die DNA kann auch *in vivo* mit der ChIP (Chromatin Immunoprecipitation)-Methode untersucht werden (Tang und Perry 2003). Der Vorteil ist, dass das Ausgangsmaterial aus der intakten Zelle entnommen werden kann. Die Protein-DNA-Interaktion wird dabei zunächst durch Zugabe von Formaldehyd fixiert, wonach die Zellen zerstört werden und die DNA in kleinere Fragmente zerlegt wird. Der Protein-DNA-Komplex wird anschließend mit einem spezifischen Antikörper isoliert. Das Protein wird danach von der DNA gelöst. Der DNA-Bereich wird anschließend durch PCR amplifiziert und sequenziert.
Die ChIP Methode ist wertvoll, um *in vivo* Daten zu erhalten. Jedoch ist sie aufwändig und steht nicht für alle Transkriptionsfaktoren zur Verfügung.

1.3 Beschreibung von Transkriptionsfaktor-Bindungsstellen

Um die Regulation der Transkription zu analysieren und zu verstehen, sind Bindungsstellen von Transkriptionsfaktoren von großer Bedeutung. Einerseits stellen diese reale Bindungsstellen dar, andererseits sind sie, wenn mehrere Bindungsstellen eines Faktors gesammelt vorliegen, geeignet, um neue putative Transkriptionsfaktor-Bindungsstellen vorherzusagen.

Am Anfang einer Sammlung von Bindungsstellen steht eine einzelne DNA-Sequenz, an die ein Transkriptionsfaktor binden kann. Je nach Art des Experiments, welches zur Ermittlung der Sequenz eingesetzt wurde, kann diese recht lang sein. Wurde kein DNAse footprint durchgeführt, ist unbekannt, welche der eingesetzten Nukleotide genau vom Faktor gebunden werden.

Werden nun mehrere DNA-Sequenzen gesammelt, an die ein einziger Transkriptionsfaktor bindet, können diese Sequenzen zusammengefasst werden. Die Bindungssequenz eines Transkriptionsfaktors enthält eine konservierte Kernsequenz, innerhalb derer ein Nukleotidaustausch in der Regel zu einer verminderten Bindungsfähigkeit des Faktors an die DNA führt oder gar zu dem völligen Unvermögen des Faktors, an diese DNA-Sequenz zu binden, sowie flankierende Nukleotide, bei denen eine größere Variabilität der Nukleotide toleriert wird. Generell ist es nun wichtig, die Kernsequenz des Faktors zu identifizieren und die gesammelten Sequenzen so auszurichten, dass die Kernsequenzen untereinander stehen.

Es gibt zwei Formen des Zusammenfassens von mehreren DNA-Bindungssequenzen. Aus den einzelnen Sequenzen kann eine einzige Konsensus-Sequenz erstellt werden, für die der IUPAC Mehrdeutigkeitscode verwendet wird. Alternativ können die Bindungssequenzen durch Matrizen dargestellt werden.

Zunächst wird näher auf die Konsensus-Sequenz eingegangen, deren IUPAC-Code in Tabelle 1 angegeben ist.

Tabelle 1: Aufschlüsselung des IUPAC-Mehrdeutigkeits Codes (aus National Center of Biotechnology Information NCBI; http://www.ncbi.nlm.nih.gov/SNP/iupac.html/)

IUPAC-Buchstabe	Mögliche(s) Nukleotid(e)
A	A
C	C
G	G
T	T
M	A oder C
R	A oder G
W	A oder T
S	C oder G
Y	C oder T
K	G oder T
V	A oder C oder G
H	A oder C oder T
D	A oder G oder T
B	C oder G oder T
N	A oder C oder G oder T

Zur Erstellung der Konsensus-Sequenz und der Festlegung, welcher Buchstabe an einer Position für die Sequenz verwendet wird, können die Regeln nach Cavener (1987) eingesetzt werden, welche im Folgenden aufgelistet werden.

Regel 1: Ein einzelnes Nukleotid wird angegeben, wenn seine Frequenz an einer Position mindestens 50 % beträgt und wenigstens zweimal so hoch ist wie das zweithäufigste Nukleotid.

Regel 2: Ein Zweideutigkeitsbuchstabe gibt an, dass zwei Nukleotide in mehr als 75 % in der zugrunde liegenden Sequenz an der bestimmten Position vorkommen und Regel 1 nicht zutrifft.

Regel 3: Tritt an einer Position ein bestimmtes Nukleotid gar nicht auf, und trifft keine der vorab beschriebenen Regeln zu, wird die Position durch einen Dreideutigkeitsbuchstaben kenntlich gemacht.

Regel 4: Alle anderen Frequenzen werden durch den Buchstaben N gekennzeichnet.

Mit einer Konsensus-Sequenz ist es möglich, genomweit nach putativen Transkriptionsfaktor-Bindungsstellen zu suchen. Aufgrund der Regeln ist jedoch ersichtlich, dass die Konsensus-Sequenz eine recht grobe Zusammenfassung der in diese Sequenz eingehenden Einzelsequenzen darstellt, und zwar genau dann, wenn nicht nur ein Nukleotid an einer Position auftritt.

Eine Alternative, mehrere DNA-Sequenzen zusammenzufassen, stellt die Erstellung von sogenannten Alignment Matrizen dar, welche in Abb. 3 dargestellt ist (Hertz und Stormo, 1999). Diese liefern eine Zusammenfassung der Ursprungssequenzen. Eine solche Matrize wird genau dann erhalten, wenn die untereinander geschriebenen Sequenzen in Spalten á 1 Nukleotid aufgeteilt werden. Anschließend wird dann die Anzahl jedes der vier vorkommenden Nukleotide A, C, G, T an einer Position (Spalte) aufsummiert und tabellarisch aufgeführt. In Abb. 3 sind beispielhaft DNA-Sequenzen aufgeführt, aus denen sowohl eine Matrize als auch die Konsensus-Sequenz erstellt wurden. Ersichtlich wird, dass eine Matrize die in sie eingehenden Sequenzen genauer widerspiegeln als Konsensus-Sequenzen.

Beispielhafte DNA-Sequenzen

```
TGACGTTT
AAACGTGG
TGACGTAC
CCACGTCC
TCACGTTC
```

Resultierende Alignment Matrize

Position / Nucleotid	1	2	3	4	5	6	7	8
A	1	1	5	0	0	0	1	0
C	1	2	0	5	0	0	1	3
G	0	2	0	0	5	0	1	1
T	3	0	0	0	0	5	2	1

Resultierende Consensus-Sequenz

```
TSACGTNC
```

Abb. 3: Möglichkeiten der Darstellung einer Sammlung von DNA-Bindungsstellen in Form einer Alignment Matrize und in der Konsensus-Sequenz Schreibweise.

In dieser Arbeit wurden für die Identifizierung von putativen Transkriptionsfaktor-Bindungsstellen Matrizen eingesetzt, da diese von Vorhersage-Programmen verarbeitet werden können und die zugrundeliegenden Sequenzen genauer widerspiegeln als Konsensus-Sequenzen (Hertz und Stormo, 1999).

1.4 *Arabidopsis thaliana* als Modellorganismus

Die Ackerschmalwand (*Arabidopsis thaliana*) ist eine Pflanze, die auch auf heimischen Wiesen zu finden ist. Sie ist in den Blickpunkt der Wissenschaft gelangt und ein Modellorganismus in der Pflanzengenetik geworden, weil sie wertvolle Eigenschaften für genetische Untersuchungen aufweist. Zum einen lässt sich die Pflanze leicht kultivieren und ist recht klein, benötigt also wenig Platz im Vergleich zu anderen Pflanzenarten, die in der Genetik verwendet werden, wie z.B. Tabak. Zum anderen hat sie eine kurze Generationszeit von etwa 6 Wochen und produziert eine Vielzahl an Samen, etwa 1000 pro Pflanze. Es ist also bei genetischen Experimenten mit schnellen Ergebnissen zu rechnen. Ein weiterer Vorteil von *Arabidopsis thaliana* ist die leichte Transformierbarkeit mittels *Agrobacterium tumefaciens*.

Die Sequenzierung des kompletten Genoms wurde von der *Arabidopsis* Genome Initiative im Jahre 2000 veröffentlicht und der wissenschaftlichen Gemeinde frei zur Verfügung gestellt (*Arabidopsis* Genome Initiative, 2000). Die genomische Annotation wurde in den darauffolgenden Jahren von TIGR (The Institute for Genomic Research) fortgeführt.
Das Genom der Pflanze ist auf fünf Chromosomen verteilt und recht dicht gepackt, etwa alle 4,5 kb ist ein Gen lokalisiert. Es ist mit etwa 119 Mb für höhere Pflanzen verhältnismäßig klein und weist wenige repetitive Sequenzen auf (Haas *et al.,* 2005). Über 5 % des Genoms kodieren für mehr als 1500 Transkriptionsfaktoren, wobei etwa 45 % dieser Faktoren nur in Pflanzen vorkommen (Riechmann *et al.,* 2000). Im Vergleich dazu weisen die Genome von *Drosophila melanogaster*, *Caenorhabditis elegans* und *Saccharomyces cerevisiae* nur zu 4,5 %, 3,5 % bzw. 3,5 % Transkriptionsfaktoren auf (Riechmann *et al.,* 2000).
Da *Arabidopsis thaliana* durchsequenziert ist, eignet sie sich gut für bioinformatische Analysen, wobei Vorhersagen aufgrund ihrer positiven Eigenschaften bei der Kultivierung und Transformation in Laborexperimenten einfach überprüft werden können.

1.5 AthaMap, eine genomweite Karte von TFBSn bei *Arabidopsis thaliana*

Das Genom von *Arabidopsis thaliana* wurde im Jahr 2000 veröffentlicht (*Arabidopsis* Genome Initiative, 2000). Daher ist *Arabidopsis* für genomweite bioinformatische Analysen gut geeignet.
Die AthaMap Datenbank wurde erstellt, um die Lokalisierung putativer und verifizierter Transkriptionsfaktor-Bindungsstellen im *Arabidopsis thaliana* Genom darzustellen (Steffens *et al.,* 2004). Die Daten sind unter http://www.athamap.de öffentlich zugänglich. Ziel ist es, dem Benutzer die Möglichkeit zu geben, auf bioinformatischem Wege die Regulation bzw. Koregulation von Genen durch einzelne Transkriptionsfaktoren oder auch durch eine mögliche Kolokalisation von Transkriptionsfaktor-Bindungsstellen an bestimmten Positionen relativ zu den entsprechenden Genen vorherzusagen. Da die Bioinformatik keine 100 %igen Aussagen treffen kann, sind diese Vorhersagen dazu einsetzbar, eine Arbeitshypothese zu erstellen und somit die Anzahl der Laborversuche bei wissenschaftlichen Fragestellungen zu minimieren.

Zur Erstellung der Datenbank, ausgehend vom TIGR Release 5 (Haas *et al.,* 2005), wurden die DNA-Sequenz sowie die Information über die Lokalisierung und Funktion der Gene aus *Arabidopsis thaliana* extrahiert. Für die Identifizierung von putativen Transkriptionsfaktor-Bindungsstellen wurden mittels Literaturrecherche Alignment Matrizen von pflanzlichen Transkriptionsfaktoren ermittelt. Die Identifizierung der entsprechenden Bindungsstellen im Genom wurde mit der Software Patser (unter Verwendung des automatisch generierten

„auto-score“) durchgeführt (Hertz und Stormo, 1999). Um die Daten zu verwalten, wurde eine Datenbankstruktur entwickelt, in die die Daten annotiert wurden. Zusätzlich zu den einzelnen Transkriptionsfaktor-Bindungsstellen wurden bekannte kombinatorische Elemente zweier Transkriptionsfaktor-Bindungsstellen, die in einem bestimmten Abstand voneinander auftreten, vorkalkuliert und ebenfalls mit in die Datenbank aufgenommen (Steffens *et al.*, 2004, Steffens *et al.*, 2005). Des Weiteren wurden einzelne DNA-Bindungsstellen von *Arabidopsis thaliana* Transkriptionsfaktoren, für die keine Alignment Matrize gebildet werden konnte, aus der Literatur entnommen, wobei in diesen Fällen die genomweite Suche nach diesen einzelnen DNA-Sequenzen mit einem Perl-Skript durchgeführt wurde. Einerseits wurde hierbei die in der Literatur vorgefundene DNA-Sequenz im *Arabidopsis thaliana* Genom eingesetzt, jedoch schloss sich zusätzlich noch eine weitere Suche mit einer entsprechend verkürzten Länge der Ursprungssequenz mit der Bedingung, dass die Kernsequenz sowie einige flankierende Nukleotide in dieser Sequenz enthalten sind, an. Die so erhaltenen Daten wurden ebenfalls in die Datenbank aufgenommen (Bülow *et al.,* 2006).

Die Datenbank wurde am Institut für Genetik der Technischen Universität Braunschweig entwickelt und steht dort auf einem Microsoft SQL Server zur Verfügung. Um die Datenbank öffentlich nutzbar zu machen, wurde ein Webinterface implementiert. Wie in Abb. 4 dargestellt hat der Benutzer die Möglichkeit, nach einer bestimmten Position im Genom oder nach einem Gen zu suchen. Im Sequenzfenster werden die entsprechende DNA-Sequenz sowie die annotierten putativen Transkriptionsfaktor-Bindungsstellen basengenau angezeigt. Informationen über den Transkriptionsfaktor können über einen Mausklick auf den Faktornamen (in blau) angezeigt werden. Links neben dem Faktornamen ist die Länge der Bindungsstelle dargestellt, wobei ein Mouse-over Fenster nähere Angaben über die genomische Lokalisierung und über die Parameter der identifizierten putativen Bindungsstelle liefert. Die DNA-Sequenzdarstellung unterscheidet zwischen intergenischen Bereichen, UTR, Exon sowie den Introns von Genen (Galuschka *et al.,* 2007).

Das AthaMap Web-Interface bietet nicht nur die Möglichkeit, nach Positionen im Genom oder nach Genen zu suchen, in deren Bereiche dann putative oder verifizierte Transkriptionsfaktor-Bindungsstellen anzeigt werden, sondern ermöglicht auch, gezielt nach Kolokalisationen von Transkriptionsfaktor-Bindungsstellen zu suchen. Im Untermenü Colocalization ist es möglich, zwei Transkriptionsfaktoren oder auch vorannotierte kombinatorische Elemente, sowie einen Mindest- und Maximalabstand der beiden Bindungsstellen auszuwählen. Das Ergebnis der Datenbankabfrage wird in Form von Tabellen ausgegeben, die absolute Positionen und die zugeordneten Gene enthalten. Diese sind durch Hyperlinks mit dem Sequenzfenster der AthaMap verlinkt, so dass auch die

Umgebung der putativen kombinatorischen Elemente visuell analysiert werden kann. Die Analyse liefert einen Eindruck über mögliche koregulierte Gene und kann einen Hinweis darauf geben, ob weitere Transkriptionsfaktoren an der Regulation der Gene beteiligt sein können.

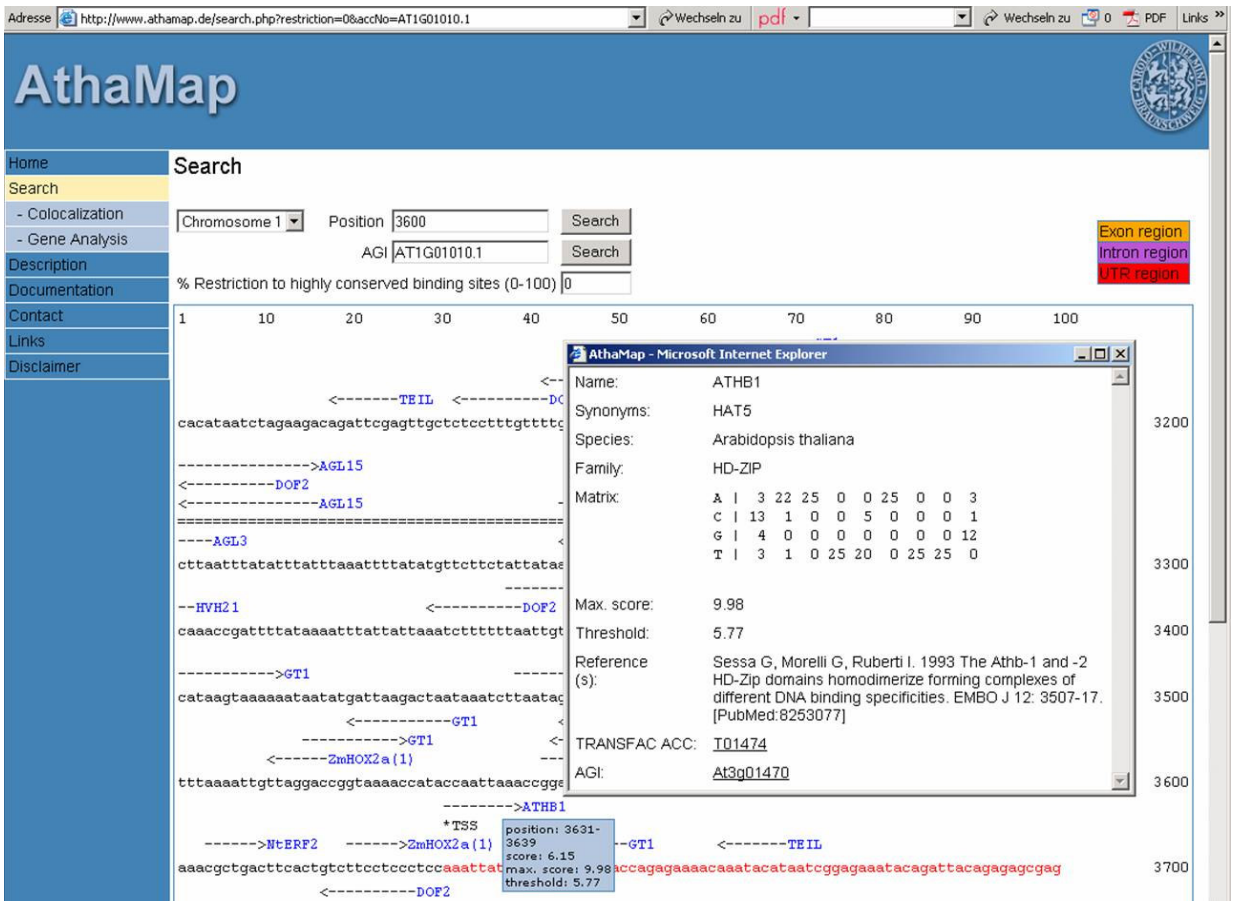
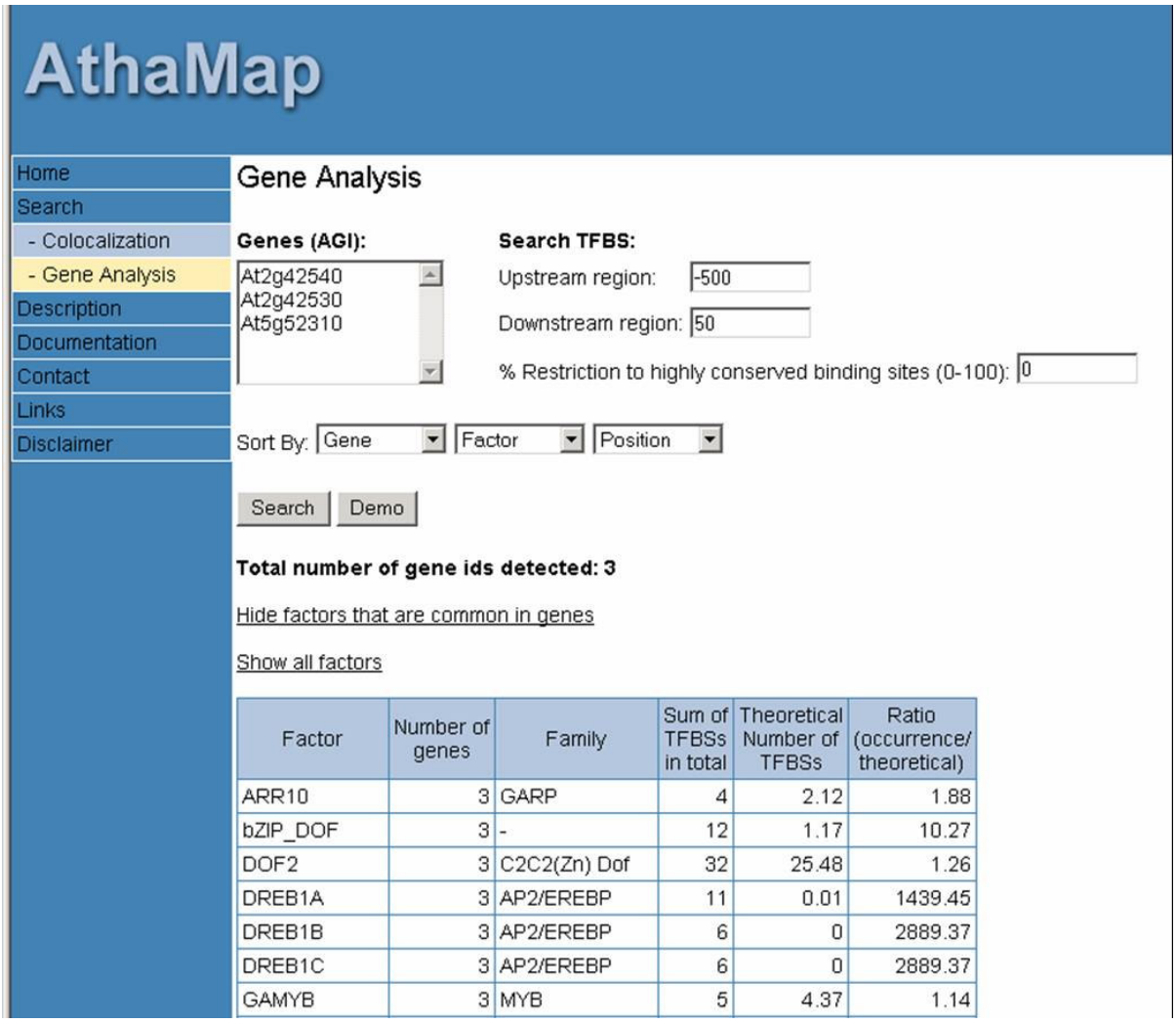


Abb. 4: Ausschnitt aus der Webseite von AthaMap. In der Navigationsleiste wurde Search ausgewählt, wonach das Gen At1g01010.1 eingegeben wurde und der Search-Button aktiviert wurde. Angegeben ist der entsprechende DNA-Sequenzabschnitt zusammen mit den putativen Transkriptionsfaktor-Bindungsstellen. Weitere Informationen über die genomische Position und einiger weiterer Merkmale der spezifischen putativen Bindungsstelle, in diesem Fall eine ATHB1 Bindungsstelle, werden durch ein Mouse-over Fenster angezeigt, wenn die Markierung der Bindungsstelle mit dem Mauszeiger berührt wird. Ein Pop-up Fenster, welches nähere Informationen über den Transkriptionsfaktor und die zugrundeliegenden Parameter, die für die Identifizierung der Bindungsstellen verwendet wurden, erscheint nach einem Mausklick auf einen Transkriptionsfaktornamen.

Mit der Gene Analysis Funktion kann, ausgehend von einer eingegebenen Genliste, nach gemeinsam auftretenden Transkriptionsfaktor-Bindungsstellen gesucht werden, um einen Hinweis darüber zu erhalten, ob die Gene möglicherweise koreguliert sind. In Abb. 5 ist das

Ergebnis einer solchen Abfrage angegeben. Da Transkriptionsfaktor-Bindungsstellen nicht nur auf den upstream-Bereich eines Gens beschränkt sind, ist es möglich, in einer benutzerdefinierten Region um den Genstart herum (Transkriptionsstartpunkt, wenn dieser bekannt ist, ansonsten dient der Translationsstartpunkt als Genstart) nach Transkriptionsfaktor-Bindungsstellen in der Datenbank zu suchen. Die Ausgabe umfasst einerseits wie bei der Kolokalisationsanalyse die einzelnen Positionen der Bindungsstellen, die mit dem Sequenzfenster der AthaMap verlinkt sind, aber auch weitere Tabellen, die es dem Benutzer ermöglichen, sich einen Überblick über Häufigkeiten der Bindungsstellen zu verschaffen und auf den ersten Blick Gemeinsamkeiten zwischen den Genen zu erkennen.



Factor	Number of genes	Family	Sum of TFBSs in total	Theoretical Number of TFBSs	Ratio (occurrence/ theoretical)
ARR10	3	GARP	4	2.12	1.88
bZIP_DOF	3	-	12	1.17	10.27
DOF2	3	C2C2(Zn) Dof	32	25.48	1.26
DREB1A	3	AP2/EREBP	11	0.01	1439.45
DREB1B	3	AP2/EREBP	6	0	2889.37
DREB1C	3	AP2/EREBP	6	0	2889.37
GAMYB	3	MYB	5	4.37	1.14

Abb. 5: Ausschnitt aus der Gene Analysis Funktionalität von AthaMap. DREB1A hat beispielsweise insgesamt 11 Bindungsstellen, die bei allen eingegebenen Genen zu finden sind. Bei angenommener Gleichverteilung der Bindungsstellen würden nur 0,01 Bindungsstellen in dem analysierten Bereich erwartet werden. Somit wurden 1436,45-mal mehr Bindungsstellen in dem Bereich gefunden als erwartet. Es existiert also eine starke Überrepräsentierung, was auf eine biologische Relevanz hindeutet.

Die tatsächliche Anzahl der Bindungsstellen wird einer theoretischen Anzahl unter Annahme einer gleichmäßigen Verteilung von Bindungsstellen im Genom gegenübergestellt. Der Quotient (Ratio) liefert einen ersten Eindruck über ein verstärktes (Ratio > 1) bzw. ein vermindertes Auftreten der Bindungsstellen (Ratio < 1). Liegt der Wert bei 1 treten so viele Bindungsstellen wie erwartet auf (Abb. 5).

Die AthaMap ist bei wissenschaftlichen Fragestellungen bezüglich der Lokalisierung von Transkriptionsfaktor-Bindungsstellen im *Arabidopsis thaliana* Genom sowie der Mechanismen der Koregulierung von Genen einsetzbar.

Für diese Arbeit wurde die Datenbankstruktur der AthaMap sowie die annotierten Daten eingesetzt, um die positionelle Verteilung von Transkriptionsfaktor-Bindungsstellen im *Arabidopsis thaliana* Genom zu analysieren.

1.6 Ziele dieser Arbeit

Im Rahmen der vorliegenden Doktorarbeit sollten Verteilungsanalysen von putativen pflanzlichen Transkriptionsfaktor-Bindungsstellen im *Arabidopsis thaliana* Genom durchgeführt werden (TIGR Release 5, Haas *et al.,* 2005). Hierbei sollte festgestellt werden, ob es Anhäufungen von Transkriptionsfaktor-Bindungsstellen im Promotor, in der 5'UTR, in Introns, in Exons sowie in der 3'UTR gibt. Die Anzahl der Introns und Exons pro Gen kann stark variieren, deshalb sollte bei detaillierten Verteilungsanalysen innerhalb von Genen eine sinnvolle Einschränkung der Anzahl der untersuchten Introns, Exons, usw. erfolgen. Des Weiteren sollte die betrachtete Länge des intergenischen Bereichs sowie innerhalb der Gene eingeschränkt werden.

Transkriptionsfaktor-Bindungsstellen sollten unter Verwendung von Matrizen und dem Einsatz der Software Patser (Hertz und Stormo, 1999) genomweit identifziert werden. Um die aus Analysen mit pflanzlichen Matrizen resultierenden Ergebnisse einschätzen zu können, sollten Hintergrundmodelle entwickelt werden.

Da putative Transkriptionsfaktor-Bindungsstellen in der Regel eine nicht zufällige Nukleotidverteilung aufweisen, sollte festgestellt werden, ob das Auftreten von Bindungsstellen an bestimmten Orten innerhalb von Genen aufgrund der Nukleotidkomposition zu erwarten ist. Daher sollte die Nukleotidverteilung im *Arabidopsis thaliana* Genom untersucht werden.

Ausgehend von den Ergebnissen sollte ein statistisches Kriterium eingesetzt werden, wodurch diese beurteilt werden können. Dadurch kann z.B. eine statistische Häufung von Transkriptionsfaktor-Bindungsstellen in bestimmten Bereichen festgestellt werden. Für einige so erhaltene Ergebnisse sollte festgestellt werden, ob die zu den Transkriptionsfaktor-Bindungsstellen gehörenden Gene möglicherweise koreguliert sind. Daher sollten Genexpressionsanalysen mit Genevestigator durchgeführt werden (Zimmermann *et al.,* 2005).

2. Material und Methoden

2.1 *Arabidopsis thaliana* Genom-Daten

Im Rahmen dieser Arbeit wurden Verteilungsanalysen von Transkriptionsfaktor-Bindungsstellen im *Arabidopsis thaliana* Genom durchgeführt. Für diese Analysen wurde die zu Beginn der Arbeit aktuelle Genomannotation eingesetzt, der TIGR Release 5 (Haas *et al.,* 2005). Frei im Internet ist dieser zugänglich unter ftp://ftp.tigr.org/pub/data/a_thaliana/ath1/PSEUDOCHROMOSOMES. Die entsprechenden Daten stehen in Form von XML-Dateien sowie einer beschreibenden DTD-Datei im Internet frei zur Verfügung. Diese Daten mussten zunächst prozessiert werden, um sie als Grundlage für Genom- und Verteilungsanalysen verwenden zu können.

Es erfolgte eine Einschränkung auf die proteinkodierenden Gene. In Abb. 6 ist die relevante Genannotation dargestellt, welche hierarchisch aufgebaut ist (Haas *et al.,* 2005). Es befinden sich jeweils die Start- und Stopppositionen bezüglich der in Abb. 6 dargestellten Bereiche in den XML-Dateien. Das längste Transkript eines Gens wird als „Transcription Unit“ (TU) bezeichnet und beinhaltet neben den positionellen Angaben auch die AGI (*Arabidopsis* Gene Identifier) Nummer, z.B. At1g01060. Jedes Gen enthält mindestens ein MODEL, welches Positionsangaben zum translatierten Bereich enthält. Weist ein Gen mehr als ein MODEL auf, stellen diese Splicevarianten des entsprechenden Gens dar. Bei mehr als einem MODEL pro Gen lauten die AGI Nummern z.B. At1g01060.1 und At1g01060.2 usw. EXONs beinhalten die positionellen Informationen zu allen transkribierten Bereichen mit Ausnahme von Introns. Die Positionsangaben bezüglich der rein translatierten Bereiche sind in den CDSs aufgelistet. Transkribierte, aber nicht translatierte, Bereiche sind in den UTRs aufgeführt. Positionsangaben bezüglich der Introns befinden sich in den XML-Dateien nicht, resultieren jedoch aus den Differenzen der annotierten Positionsangaben von Exons. Zusätzlich zu der Annotationsdarstellung in Abb. 6 können auch die UTRs von Genen Introns beinhalten.

Abb. 6 : Prinzipieller Aufbau der Annotation für proteinkodierende Gene aus der tigrxml.dtd-Datei (TIGR Release 5)

Aus den XML-Dateien wurden mittels des selbst verfassten Perl-Skripts CGparseTIGR.pl die kompletten DNA-Sequenzen der fünf Chromosomen sowie die positionellen Informationen über die Genbereiche extrahiert und die Positionen der Introns berechnet. Extrahiert wurden nur die proteinkodierenden Gene. Hierzu werden laut TIGR Release 5 (Haas *et al.,* 2005) ebenfalls Pseudogene und Transposons gezählt. Die durch Einsatz des Perl-Skripts gewonnenen Informationen wurden in Tabellenform gespeichert, wobei für die Positionsangaben eines einzelnen Bereichs (TU, MODEL, EXON, CDS, UTR bzw. INTRON) jeweils eine Zeile resultierte.

Zu Beginn der Arbeit waren die DNA-Sequenz und die Geninformationen des TIGR Release 4 in der bereits bestehenden Datenbank TFBSites (Kapitel 2.5.3) implementiert. Um eine Kompatibilität mit der vorhandenen Datenbankstruktur zu erzielen, war eine weitere Prozessierung der Genanntoationsdateien des TIGR Release 5 notwendig. Diese erfolgte durch die Verwendung der selbst verfassten Perl-Skripte CGCleanForImport.pl und CGJoinFilesAddID.pl. Anschließend stand eine einzige Datei, die die gesamte Genannotation der fünf Chromosomen vereinigt, zur Verfügung. Durch Verwendung des selbst verfassten Perl-Skriptes SelectModels.pl wurden sämtliche MODEL-Zeilen extrahiert, aus denen dann die Datenbanktabelle cds der TFBSites erzeugt wurde (Kapitel 2.5.3). Parallel dazu erfolgte mittels des selbst erstellten Skriptes CGFeatureUpDown.pl eine Durchnummerierung der EXONs, CDSs, UTRs sowie der Introns. Diese wurde in 5'-3'-Richtung der jeweiligen Gene vorgenommen. Aus den Daten wurde die Datenbanktabelle cds_all generiert. Diese Tabelle ist in Kapitel 2.5.3 näher erläutert.

Für sämtliche nachfolgenden Analysen wurde pro Gen nur ein MODEL, also nur eine Splicevariante, sowie je nach Bedarf dessen EXON(s), CDS(s), UTR(s) und INTRON(s) eingesetzt. 2189 Gene weisen mindestens eine Splicevariante auf. Damit es nicht zu einer Überrepräsentierung von einzelnen Genen und zu einer Überlappung von Splicevarianten kommt, wurde das MODEL mit der niedrigsten AGI-Nummer ausgewählt (Beispiel: bei den Splicevarianten At1g01060.1 und At1g01060.2 das MODEL mit der Nummer At1g01060.1). Insgesamt wurden 29993 Gene bei den folgenden Analysen mit einbezogen. Zu diesen Genen zählen 3786 Pseudogene bzw. Transposons. Bei 16319 Genen (54,41 %) ist der Transkriptionsstartpunkt bekannt.

2.1.1 Festlegung der zu analysierenden Genbereiche

Die Anzahl der Introns und Exons usw. in Abb. 6 können von Gen zu Gen stark variieren. Um einen Überblick über die Anzahl der Bereiche pro Gen zu erhalten, wurden, ausgehend von der Datenbanktabelle cds_all (Kapitel 2.5.3), mit dem MS Query Analyser Datenbankabfragen durchgeführt. Wie in Tabelle 2 aufgeführt ist, unterscheiden sich z.B. die maximale Anzahl der Introns und Exons und die durchschnittliche Anzahl dieser Bereiche erheblich voneinander. Für die weiteren Analysen erfolgte eine Beschränkung auf die durchschnittliche Anzahl der jeweiligen Bereiche pro Gen, wobei nach unten abgerundet wurde. Es werden daher nur die erste 5'UTR, die ersten fünf Introns, die ersten vier Exons und die erste 3'UTR berücksichtigt. Das Auftreten von mehr Introns als Exons ist in der Tatsache begründet, dass die 5'UTR und die 3'UTR ebenfalls Introns beinhalten können. Obwohl normalerweise ein Gen nur eine 5'UTR enthält, werden die durch Introns unterbrochenen 5'UTRs separat angegeben, was zu einer Häufung von bis zu 12 5'UTRs für für ein Gen führen kann. Ähnliches gilt für die 3'UTR.

Tabelle 2: Maximal auftretende Anzahl von 5'UTRs, Introns, Exons und 3'UTRs pro Gen sowie die durchschnittliche Anzahl der Bereiche, bezogen auf alle 29993 Gene.

Bereich	Maximale Anzahl in einem Gen	Durchschnittliche Anzahl pro Gen
5'UTR	12	1,19
Intron	78	5,48
Exon	73	4,72
3'UTR	20	1,05

Danach wurden, ebenfalls durch Datenbankabfragen, die Nukleotidlängen der 5'UTR, Introns, Exons und der 3'UTR sowie die durchschnittliche Länge der Gene bestimmt. Ein Gen ist im Durchschnitt 2246 Nukleotide lang, wobei das längste Gen 31257 Nukleotide und das kürzeste 74 Nukleotide umfasst. Wie aus Tabelle 3 ersichtlich, unterscheiden sich die maximale Länge der 5'UTR, der Introns, der Exons sowie der 3'UTR und die durchschnittliche Länge dieser Bereiche ebenfalls stark voneinander.

Tabelle 3: Maximale Länge der zu analysierenden Bereiche der Gene aus Tabelle 2 sowie die durchschnittliche Länge der jeweiligen Bereiche, angegeben in Basenpaaren.

Bereich	Nummer	Max. Länge des Bereichs [bp]	Durchschnittliche Länge [bp]
5UTR	1	2331	119
Intron	1	7884	248
	2	4947	176
	3	4162	153
	4	3122	141
	5	3480	136

Exon	1	15195	692
	2	5538	259
	3	5966	219
	4	7713	192
3UTR	1	3118	223

Da die Datenbanktabelle cds_all ebenfalls die intergenischen Bereiche enthält (Kapitel 2.5.2), konnte diese durchschnittliche Nukleotidlänge ebenfalls mittels Datenbankabfragen ermittelt werden. Der intergenische Bereich beträgt im Durchschnitt 1793 bp (Maximallänge = 72791 bp).

2.1.2 Analyse der Nukleotidkomposition von DNA-Sequenzen

Zur Analyse der Nukleotidzusammensetzung der Genbereiche aus Kapitel 2.1.1, also die erste 5'UTR, die ersten fünf Introns, die ersten vier Exons und die erste 3'UTR, wurden die Positionsangaben aus der Datenbanktabelle cds_all (Kapitel 2.5.3) verwendet. Sie wurden mittels SQL-Skripten abgefragt. Das jeweilige Ergebnis wurde als Textdatei gespeichert. Zusätzlich zu den in Kapitel 2.1.1 ermittelten Genbereichen wurden auch die am höchsten nummerierte 5'UTR, das höchste Intron, Exon sowie die höchste 3'UTR der entsprechenden Gene auf ihre Nukleotidkomposition hin analysiert. Wenn mehrere 5'UTRs auftreten, wird die 5'UTR durch Introns unterbrochen. Die höchste 5'UTR ist dann diejenige, die dem Translationsstart am nächsten liegt. Im Falle der Introns, Exons und der 3'UTR liegen die am höchsten nummerierten am weitesten vom Transkriptionsstartpunkt entfernt. Diese Positionsangaben wurden mit den TFBSitesTools (Kapitel 2.5.3) extrahiert.

Mit dem selbst verfassten Perl-Skript ExtractSequencesChooseReverseFeature.pl und der fünf Chromosomensequenzdateien (Kapitel 2.1) erfolgte die Extraktion der entsprechenden DNA-Sequenzen sowie die Berechnung der Nukleotidkomposition aus der jeweiligen Textdatei basierend auf den Datenbankabfragen der Datenbanktabelle cds_all. Um eine grafische Darstellung der DNA-Sequenzen zu ermöglichen, für die Sequenzlogos (Kapitel 2.8) eingesetzt wurden, können mit dem Perl-Skript zusätzlich auch DNA-Sequenzen einer bestimmten Länge extrahiert werden, um mittels WebLogo direkt das entsprechende Sequenzlogo berechnen lassen zu können (Kapitel 2.8).
Eine weitere Option des Perl-Skriptes ist der variable Bezugspunkt. Die DNA-Sequenzen sowie die Nukletotidkompositionen können vom Anfang oder vom Ende des Bereichs ausgerichtet und ausgegeben werden.

2.2 Matrizenauswahl

Für die Verteilungsanalysen wurden Alignment Matrizen von pflanzlichen sowie nichtpflanzlichen Transkriptionsfaktoren eingesetzt. Auf die Auswahl wird im Folgenden eingegangen. Anschließend wurde für jede Matrize eine entsprechende Zufallsmatrize generiert.

2.2.1 Pflanzliche Matrizen

Für die Detektion von putativen Transkriptionsfaktor-Bindungsstellen im *Arabidopsis thaliana* Genom wurden Alignment Matrizen eingesetzt, die Bindungsstellen eines Transkriptionsfaktors zusammenfassen (Kapitel 1.3). Um durch eine anschließende genomweite Suche nach putativen TFBSn mit der Software Patser (Kapitel 2.4) potentiell funktionale Bindungsstellen erhalten zu können, wurden pflanzliche Matrizen ausgewählt (Tabelle 4). Diese Matrizen wurden aus Publikationen extrahiert und entsprechen denen aus der AthaMap Publikation aus dem Jahr 2005 von Steffens *et al.* Die Transkriptionsfaktoren stammen aus *Arabidopsis thaliana* und aus anderen Pflanzenspezies. Die verwendeten DNA-Sequenzen wurden in der Regel experimentell identifiziert. Im Falle von TBP und CBF wurde aus Promotorsequenzen mehrerer Spezies je eine Alignment Matrize erstellt (Tabelle 4). Insgesamt wurden 42 Alignment Matrizen für 36 pflanzliche Transkriptionsfaktoren aus 15 bekannten Klassen (plus TBP und CBF) für die Analysen eingesetzt. Aufgrund früherer Analysen weisen Matrizen derselben Klasse untereinander wenige Redundanzen auf. Die Vorgehensweise einer Redundanzanalyse ist in Steffens *et al.,* 2004 beschrieben.

Tabelle 4: Liste der pflanzlichen Transkriptionsfaktoren, für die eine Alignment Matrize in dieser Arbeit verwendet wurde. Wurde für einen Faktor mehr als eine Matrize eingesetzt, ist dies durch eine eingeklammerte Zahl hinter dem Faktornamen gekennzeichnet. In einigen Fällen ist durch „_M…“ am Ende des Faktors die ID der identischen Matrize, annotiert bei TRANSFAC©, angegeben. Des Weiteren sind die Klasse des Transkriptionsfaktors, die Spezies, aus der der Faktor stammt, sowie die Referenz, aus der die Alignment Matrize extrahiert wurde, angegeben. Die Matrizen selbst sind im Anhang in Tabelle 31 aufgeführt.

Transkriptionsfaktor	Klasse	Spezies	Referenz
ABF1	bZIP	*Arabidopsis thaliana*	Choi *et al.,* 2000
ABI4(1)	AP2/EREBP	*Zea mays*	Niu *et al.,* 2002
ABI4(2)	AP2/EREBP	*Zea mays*	Niu *et al.,* 2002
AG_M00151	MADS	*Arabidopsis thaliana*	Huang *et al.,* 1993
AGL15	MADS	*Arabidopsis thaliana*	Tang und Perry, 2003

AGL3	MADS	*Arabidopsis thaliana*	Huang et al, 1995
AGP1	GATA	*Nicotiana tabacum*	Sugimoto *et al.,* 2003
ALFIN1	HD-PHD	*Medicago sativa*	Bastola *et al.,* 1998
ANT	AP2/EREBP	*Arabidopsis thaliana*	Nole-Wilson und Krizek, 2000
ATHB1	HD-ZIP	*Arabidopsis thaliana*	Sessa *et al.,* 1993
ATHB5	HD-ZIP	*Arabidopsis thaliana*	Johannesson *et al.,* 2001
ATHB9_M00417	HD-ZIP	*Arabidopsis thaliana*	Sessa *et al.,* 1998
AtMYB15	MYB	*Arabidopsis thaliana*	Romero *et al.,* 1998
AtMYB77	MYB	*Arabidopsis thaliana*	Romero *et al.,* 1998
AtMYB84	MYB	*Arabidopsis thaliana*	Romero *et al.,* 1998
bZIP910(1)	bZIP	*Antirrhinum majus*	Martinez-Garcia *et al.,* 1998
bZIP910(2)	bZIP	*Antirrhinum majus*	Martinez-Garcia *et al.,* 1998
bZIP911(1)	bZIP	*Antirrhinum majus*	Martinez-Garcia *et al.,* 1998
bZIP911(2)	bZIP	*Antirrhinum majus*	Martinez-Garcia *et al.,* 1998
CBF	-	*diverse Spezies*	Shahmuradov *et al.,* 2003
CDC5_M00361	MYB	*Arabidopsis thaliana*	Hirayama und Shinozaki, 1996
DOF2_M00353	DOF	*Zea mays*	Yanagisawa und Schmidt, 1999
GAMYB	MYB	*Hordeum vulgare*	Gubler *et al.,*1999
GT1	Trihelix	*Nicotiana tabacum*	Ouwerkerk *et al.,* 1999
HAHB4	HD-ZIP	*Helianthus annus*	Palena *et al.,* 1999
HVH21	HD-Knotted	*Hordeum vulgare*	Krusell *et al.,* 1997
ID1	C2H2 (Zn)	*Zea mays*	Kozaki *et al.,* 2004
MYB.PH3_M00218	MYB	*Petunia hybrida*	Solano *et al.,* 1995
MYB.PH3_M00219	MYB	*Petunia hybrida*	Solano *et al.,* 1995
O2	bZIP	*Zea mays*	Izawa et al., 1993
P	MYB	*Zea mays*	Grotewold *et al.,*

			1994
PCF2	TCP	*Oryza sativa*	Kosugi und Ohashi, 2002
PCF5	TCP	*Oryza sativa*	Kosugi und Ohashi, 2002
PIF3_M00434	bHLH	*Arabidopsis thaliana*	Martinez-Garcia *et al.,* 2000
PIF3(2)	bHLH	*Arabidopsis thaliana*	Martinez-Garcia *et al.,* 2000
RAV1_M00343	AP2/EREBP	*Arabidopsis thaliana*	Kagaya *et al.,* 1999
RAV1_M00344	AP2/EREBP	*Arabidopsis thaliana*	Kagaya *et al.,* 1999
TBP	-	*diverse Spezies*	Shahmuradov *et al.,* 2003
TEIL	AP2/EREBP	*Nicotiana tabacum*	Kosugi und Ohashi, 2000
TGA1	bZIP	*Arabidopsis thaliana*	Schindler *et al.,* 1992
TGA1a	bZIP	*Nicotiana tabacum*	Izawa *et al.,* 1993
ZAP1	WRKY	*Arabidopsis thaliana*	de Pater *et al.,* 1996

2.2.2 Nichtpflanzliche Matrizen

Mit pflanzlichen Matrizen sollten putativ funktionale Transkriptionsfaktor-Bindungsstellen in einem pflanzlichen Genom detektierbar sein. Um jedoch eine Vergleichsgruppe zu erhalten, die keine funktionalen Bindungsstellen im Genom zu erwarten verspricht, wurden aus der öffentlich zugänglichen Version von TRANSFAC© (http://www.gene-regulation.com), einer online zur Verfügung stehenden Ressource für Informationen über Transkriptionsfaktoren (Matys *et al.,* 2003), nichtpflanzliche Matrizen unter der Bedingung extrahiert, dass deren Transkriptionsfaktorklassen nicht in Pflanzen vorkommen.

Zunächst wurden 41 Matrizen extrahiert. Danach wurde mit Patser (Kapitel 2.4) genomweit nach putativen Transkriptionsfaktor-Bindungsstellen in *Arabidopsis thaliana* gesucht, das Resultat in die Datenbank TFBSites importiert und eine Redundanzanalyse durchgeführt. Ziel der Redundanzanalyse war, eine Auswahl von Matrizen zu erhalten, die möglichst wenige gleiche Transkriptionsfaktor-Bindungsstellen untereinander aufweisen. Ausgehend von der geringsten Redundanz untereinander (Daten nicht gezeigt) wurden 27 Alignment Matrizen für 24 Transkriptionsfaktoren aus 15 bekannten Klassen für die weiteren Analysen ausgewählt, welche in Tabelle 5 aufgeführt sind.

Tabelle 5: Liste der nichtpflanzlichen Transkriptionsfaktoren, für die eine Alignment Matrize in dieser Arbeit verwendet wurde. Die Alignment Matrizen wurden aus der öffentlichen Version von TRANSFAC© extrahiert. Die zu dem Transkriptionsfaktor gehörende Matrizen-ID aus TRANSFAC©, sowie die Klasse, zu der der Faktor gehört und die Spezies, aus dem der Transkriptionsfaktor stammt, sind ebenfalls angegeben. Die dazugehörigen Matrizen sind im Anhang in Tabelle 32 angeführt.

Transkriptionsfaktor	**TRANSFAC© Matrizen-ID**	**Klasse**	**Spezies**
abaA	M00027	TEA	*Aspergillus nidulans, Emericella nidulans*
AP-2alpha	M00189	bHSH	*Mus musculus*
AREB6	M00414	CH+homeo	*Homo sapiens*
AREB6	M00415	CH+homeo	*Homo sapiens*
Brachyury	M00150	T- Box	*Mus musculus*
c-Ets-1	M00032	ETS	*Mus musculus*
c-Ets-154	M00074	ETS	*Gallus gallus*
COUP	M00158	CC (rec)	*Rattus norvegicus*
CP2	M00072	Grainyhead	*Mus musculus*
Dl	M00120	Rel	*Drosophila melanogaster*
E2F	M00024	Fork head	*Homo sapiens*
E74A	M00016	ETS	*Drosophila melanogaster*
FOXA4a	M00267	Fork head	*Xenopus laevis*
GAL4	M00049	C6	*Saccharomyces cerevisiae*
GCM	M00270	GCM	*Drosophila melanogaster*
HAP3	M00288	histone fold	*Saccharomyces cerevisiae*
LAC9	M00207	C6	*Kluyveromyces lactis*
NF-kappaB1	M00194	Rel	*Rattus norvegicus*
Pax-2	M00098	paired	*Homo sapiens*
Pax-3	M00360	paired-homeo	*Mus musculus*
Pax-4a	M00373	paired homeo	*Mus musculus*
Pax-5	M00143	paired	*Mus musculus*
Pax-5	M00144	paired	*Homo sapiens*
Pax-6	M00097	paired-homeo	*Mus musculus*
POU2F1	M00136	POU	*Homo sapiens*

POU2F1	M00162	POU	*Homo sapiens*
PPAR-alpha	M00242	CC (rec)	*Rattus norvegicus*

2.2.3 Zufallsmatrizen

Für die Verteilungsanalyse von Transkriptionsfaktor-Bindungsstellen wurden pflanzliche Matrizen eingesetzt, die in Kapitel 2.2.1 angegeben sind. Diese Matrizen fassen funktionelle Bindungsstellen von pflanzlichen Transkriptionsfaktoren zusammen. Daher wird erwartet, dass bei einer genomweiten Suche nach Transkriptionsfaktor-Bindungsstellen putativ funktionelle Bindungsstellen detektiert werden können. Transkriptionsfaktor-Bindungsstellen beschreibende Matrizen haben nicht nur eine konservierte Kernsequenz, sondern auch einen bestimmten AT-Gehalt. Somit ist es möglich, dass es zu einer Anreicherung von Bindungsstellen in bestimmten Bereichen, die einen entsprechenden AT-Gehalt aufweisen, kommen kann. Um diesen Einfluss zu untersuchen, wurde für jede der 42 pflanzlichen Matrizen eine entsprechende Zufallsmatrize mit dem Visual Basic Programm SeqToMatrix erzeugt (Kapitel 2.5.1). Hierbei wurden die Anzahl der in die pflanzliche Matrize eingehenden Sequenzen und die Länge der Ursprungsmatrize beibehalten. Der AT-Gehalt der zu erzeugenden Matrize wurde geradzahlig auf den Wert der Ursprungsmatrize gerundet. Die erzeugten Zufallsmatrizen verfügen nicht über eine konservierte Kernsequenz und sind in Tabelle 33 im Anhang aufgelistet.

Für weitere Analysen wurden auch nichtpflanzliche Matrizen ausgewählt (Kapitel 2.2.2). Diese weisen, wie die pflanzlichen Matrizen, eine konservierte Kernsequenz auf. Da jedoch die dazugehörigen Faktoren nicht in Pflanzen vorkommen und auch deren Klassen in Pflanzen nicht bekannt sind, sollten bei einer Suche nach putativen Transkriptionsfaktor-Bindungsstellen keine funktionalen Bindungsstellen detektiert werden können. Jedoch wurde auch bei diesen Matrizen der prinzipielle Einfluss des AT-Gehalts auf die Verteilung von Bindungsstellen untersucht. Daher wurde auch für diese Matrizen, analog zu denen für die pflanzlichen Matrizen, je eine Zufallsmatrize unter denselben Voraussetzungen erzeugt. Die Zufallsmatrizen der nichtpflanzlichen Matrizen sind in Tabelle 34 im Anhang aufgeführt.

2.3 Zufallspositionen

Hintergrundwerte ohne Berücksichtigung einer zugrundeliegenden Matrize wurden durch Zufallspositionen, generiert mit den TFBSitesTools (Kapitel 2.5.3), ermittelt. Insgesamt wurden 315000 (42 x 7500) Zufallspositionen im Genom erzeugt. Diese Analyse der

Verteilung dieser Positionen lässt Rückschlüsse auf bestimmte prinzipielle Anreicherungen von Positionen relativ zu speziellen Bereichen bzw. Distanzen im Genom zu.

Die ausgegebenen putativen Bindungsstellen bzw. Zufallspositionen konnten mit den TFBsitesTools (Kapitel 2.5.3) in die Datenbank TFBSites importiert werden.

2.4 Matrizenbasierte Suche nach putativen TFBSn

Die Linux-basierte Software Patser (Hertz und Stormo, 1999) ermöglicht es, mit Alignment Matrizen genomweit nach putativen Transkriptionsfaktor-Bindungsstellen zu suchen. Das Programm ist online unter http://rsat.ulb.ac.be/rsat verfügbar und wurde lokal auf einem Linux-Server installiert.

Beispielhaft ist in Abb. 7 eine Alignment Matrize dargestellt. Die Werte in den Alignment Matrizen sind von der Anzahl der in die Matrize eingehenden Sequenzen abhängig. Daher werden sie zunächst von Patser umgewandelt, wodurch eine „Positional Weight Matrix“ (PWM) entsteht (Formel 1). In die Formel gehen die Anzahl der Sequenzen der Matrize, die tatsächliche Häufigkeit eines Nukleotids sowie deren beobachtete Häufigkeit an einer bestimmten Position und die erwartete Wahrscheinlichkeit für das Auftreten dieses Nukleotids mit ein. Die *a priori* Wahrscheinlichkeiten beträgt bei *Arabidopsis* thaliana: p_i (A), p_i (T) = 0,32; p_i (C), p_i (G) = 0,18 und ergibt sich aus der Nukleotidzusammensetzung des gesamten Genoms. Durch Aufsummieren der Einzelpositionen der PWM wird der Score berechnet (Formel 2). Für eine verbesserte Anschaulichkeit wird ein Beispiel angeführt (Abb. 7).

Formel 1: Berechnung der Positional Weight Matrix

$$PWM_{i,j} = \ln \frac{\left(n_{i,j} + p_i\right)/\left(N+1\right)}{p_i} \approx \ln \frac{f_{i,j}}{p_i} \qquad \text{mit } f_{i,j} = \frac{n_{i,j}}{N}$$

mit

N = Anzahl der Sequenzen, die in die Matrize eingehen

p_i = *a priori* Wahrscheinlichkeit des Nukleotids i an jeder Position

$n_{i,j}$= tatsächliche Häufigkeit des Nukleotids i an der Position j

$f_{i,j}$= Frequenz des auftretenden Nukleotids i an der Position j

Formel 2: Berechnung des Scores aus der PWM in Formel 1

$$Score = \sum_{i=1}^{n} PWM_{i,j} \text{ mit } \mathrm{n} = \text{Anzahl der Positionen der } \mathrm{PWM_{i,j}}$$

Alignment Matrize

A|	5	2	**19**	0	**19**	0	0	0	**14**	4
C|	5	10	0	0	0	0	**19**	**4**	0	5
G|	1	3	0	**19**	0	0	0	0	**3**	5
T|	8	4	0	0	0	**19**	0	**15**	**2**	5

Resultierende PWM

A|	-0,13	-0,96	**1,16**	**-2,94**	**1,16**	**-2,94**	**-2,94**	**-2,94**	**0,86**	-0,34
C|	0,42	1,09	**-2,94**	**-2,94**	**-2,94**	**-2,94**	**1,72**	**0,20**	**-2,94**	0,42
G|	-1,06	-0,07	**-2,94**	**1,72**	**-2,94**	**-2,94**	**-2,94**	**-2,94**	**-0,07**	0,42
T|	0,31	-0,34	**-2,94**	**-2,94**	**-2,94**	**1,16**	**-2,94**	**0,92**	**-0,96**	-0,13

minimaler Score

-1,06+(-0,96)+(**-2,94**)+(**-2,94**)+(**-2,94**)+(**-2,94**)+(**-2,94**)+(**-2,94**)
+(**-2,94**)+(-0,34)
= -22,94

maximaler Score

0,42+1,09+**1,16+1,72+1,16+1,16+1,72+0,92+0,86**+0,42
= 10,63

Abb. 7: Darstellung einer beispielhaften Alignment Matrize, berechneter PWM und den beiden Extremwerten der Scores.

Ausgehend von der PWM resultiert ein Score, wenn von jeder Position ein Wert in der Spalte ausgewählt wird und dann die Summe dieser Werte gebildet wird. Auf den ersten Blick wird ersichtlich, dass es zwei Maximalwerte geben muss, den minimalen Score, der die Summe der minimalen Werte einer jeden Position zusammenfasst und den maximalen Score, der die positivsten Werte jeder Position widerspiegelt. Die Extremwerte des Scores hängen von der Länge der Alignment Matrize, vom Konservierungsgrad der einzelnen Positionen in der Matrize und von der Anzahl der in die Berechnung der PWM eingehenden Sequenzen ab. Die PWM wird von Patser verwendet, um in einer genomischen Sequenz nach Übereinstimmungen mit der PWM zu suchen. Je größer die Übereinstimmung, umso größer ist auch der Score des DNA-Sequenzabschnitts, der dieselbe Länge wie die Alignment Matrize aufweist.

Patser berechnet einen Score, den eine putative Transkriptionsfaktor-Bindungsstelle aufweisen muss, um als solche ausgegeben zu werden, selbständig und beträgt für die in Abb. 7 angegebene Beispielmatrize 6,87. Dieser automatisch berechnete Score hängt unter anderem vom Informationsgehalt der Alignment Matrize ab. Je größer dieser ist, also je unwahrscheinlicher es ist, dass eine bestimmte DNA-Sequenz im Genom vorkommt, umso

selektiver sind die ausgegebenen putativen Transkriptionsfaktor-Bindungsstellen. Alignment Matrizen, die auf Zufallssequenzen beruhen, haben einen Informationsgehalt, der nahe Null liegt, was zu unverhältnismäßig vielen ausgegebenen Bindungsstellen führen würde.

Alternativ bietet Patser die Funktionalität an, dass der Benutzer einen Mindest-Score vorgegeben kann. Die ausgegebenen putativen Transkriptionsfaktor-Bindungsstellen weisen in dem Fall einen Wert zwischen dem Mindest-Score und dem maximal erreichbaren Score auf.

Um das Programm zu verwenden, ist eine Linux-Shell erforderlich. Für die Detektion von putativen Transkriptionsfaktor-Bindungsstellen wurden die folgenden Kommandozeilen verwendet:

1. automatisch berechneter Score: patser-v3d -A a:t 0.32 c:g 0.18 -m Matrizenname -f Sequenzname -c -li -d2

2. benutzerdefinierter Score: patser-v3d -A a:t 0.32 c:g 0.18 -m Matrizenname -f Sequenzname -c -ls score -M score -d2

Wurde eine putative Bindungsstelle identifiziert, bezieht sich die Positionsangabe immer auf die erste aufgetretene Position auf dem Chromosom. Ausgegeben wird die Positionsangabe auf dem annotierten DNA-Strang. Wenn sie auf dem komplementären DNA-Strang lokalisiert ist, wird die Positionsangabe durch ein C ergänzt. Liegt die putative Bindungsstelle auf dem annotierten DNA-Strang, handelt es sich bei der Positionsangabe um die erste Position der Bindungsstelle, liegt sie hingegen auf dem nicht annotierten DNA-Strang, identifiziert die Positionsangabe die letzte Position.

2.4.1 Einschränkung auf die 7500 konserviertesten putativen TFBSn

Ziel ist es, eine Verteilungsanalyse von Transkriptionsfaktor-Bindungsstellen unter Einsatz von Matrizen im *Arabidopsis thaliana* Genom durchzuführen. Mit der Software Patser kann ein automatisch generierter Score eingesetzt werden, der putative Bindungsstellen liefert. In den automatisch generierten Score geht der Informationsgehalt der Matrize mit ein. Da bei weitergehenden Analysen auch Zufallsmatrizen eingesetzt wurden, deren Informationsgehalt nahe Null liegt, kann dieser für diese Kontrollgruppen nicht eingesetzt werden. Des Weiteren wird bei dem Einsatz des automatisch generierten Scores eine stark unterschiedliche Anzahl von Bindungsstellen für die verschiedenen pflanzlichen Matrizen ausgegeben. Um die

Ergebnisse jedoch vergleichbar zu machen, wurde dieser Score benutzerdefiniert eingeschränkt, so dass approximativ die 7500 konserviertesten Bindungsstellen pro Matrize ausgegeben werden. Vorabanalysen haben gezeigt, dass bei Verwendung der 7500 ausgegebenen TFBSn pro Matrize die Kernsequenz bei diesen vorwiegend erhalten blieb (pers. Mitteilung Nils Ole Steffens). Um jedoch zu den approximativ 7500 konserviertesten Bindungsstellen zu gelangen, musste bei einigen pflanzlichen Matrizen der automatisch generierte Score abgesenkt werden.

Ein Qualitätskriterium bei den mit Matrizen und Patser detektierten putativen Transkriptionsfaktor-Bindungsstellen ist das Vorhandensein der Kernsequenz der ausgegebenen Bindungsstellen. Um festzustellen, ob es durch das manuelle Einstellen des Scores zu Qualitätseinbußen bei den ausgegebenen Transkriptionsfaktor-Bindungsstellen kommt, wurde zweimal eine genomweite Suche nach putativen Bindungsstellen vorgenommen. Zum einen wurde für jede der 42 pflanzlichen Matrizen (Kapitel 2.2.1) der automatisch generierte Score verwendet, zum anderen wurde die Anzahl der ausgegebenen putativen Transkriptionsfaktor-Bindungsstellen auf approximativ 7500 durch Anpassen des Scores eingeschränkt. Anschließend erfolgte der Import in die Datenbank TFBSites mit einer anschließenden Redundanzanalyse, um doppelte Bindungsstellen einer Matrize an ein und derselben genomischen Position auszuschließen. Die Summe der nicht redundanten Transkriptionsfaktor-Bindungsstellen für beide Patser-Einstellungen sind im Anhang in Tabelle 35 zusammengefasst. Nach der Redundanzanalyse wurde für die beiden Datensätze jeder Matrize die Kernsequenz mit den TFBSitesTools eingelesen. Als Kernsequenz wurde diejenige Sequenz bezeichnet, die bei den ermittelten Sequenzen am konserviertesten war. Dies musste nicht in allen Fällen die komplette Kernsequenz der Transkriptionsfaktor-Klasse sein. Die ermittelten Kernsequenzen befinden sich im Anhang in Tabelle 36. Nach diesem Einlesen wurde die Anzahl der putativen Transkriptionsfaktor-Bindungsstellen erfasst, welche die Kernsequenz aufwiesen. Die Auswertung ergab, dass, bei der Einstellung des Scores auf die approximativ 7500 konserviertesten Bindungsstellen, bei 24 Matrizen die Kernsequenz zu jeweils mindestens 90 % erhalten bleibt (Abb. 8). Im Falle des automatisch generierten Scores wurde bei 23 Matrizen das Vorhandensein der Kernsequenzen bei mindestens 90 % der detektierten putativen Bindungsstellen festgestellt (Daten nicht gezeigt). Es wurden also vergleichbare Ergebnisse erhalten. Interessant ist, dass bei 17 Matrizen alle detektierten putativen Bindungsstellen die Kernsequenz aufweisen (Abb. 8), wenn der Score auf 7500 Bindungsstellen pro Matrize eingestellt wird, während es nur bei 8 Matrizen zu diesem Ergebnis kommt, wenn der automatisch generierte Score eingesetzt wird. Unter Einsatz eines benutzerdefinierten Scores kann also die Anzahl der Kernsequenzen in einigen Fällen erhöht werden. Im Folgenden wurde für jede Matrize

erfasst, ob es bei der benutzerdefinierten Einschränkung des Scores zu einer Erhöhung der Anzahl der vorhandenen Kernsequenzen der Bindungsstellen kommt, ob keine Änderung auftritt oder ob die Anzahl der vorhandenen Kernsequenzen abnimmt. Dies ist in Abb. 8 farblich kenntlich gemacht. Hierbei wurde eine generelle Zusammenfassung in 10 % Schritten durchgeführt, mit Ausnahme derjenigen, deren putativen Bindungsstellen die Kernsequenz zu 100 % aufweisen.

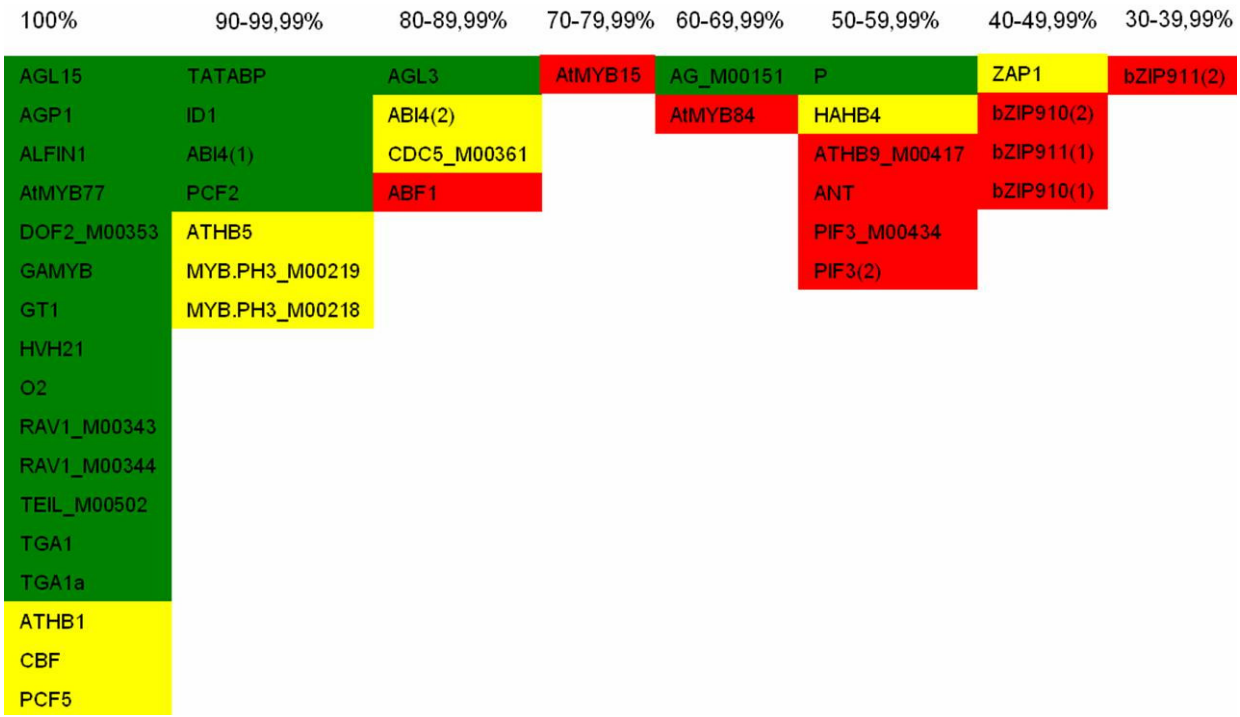


Abb. 8: Namen der pflanzlichen Matrizen, deren approximativ 7500 detektierten putativen TFBSn die Kernsequenz in dem oben angegebenen Prozentbereich aufweisen. Der Farbcode stellt einen Vergleich zwischen den approximativ 7500 konserviertesten TFBSn und den Bindungsstellen, detektiert mit dem auto-Score von Patser, dar. „Grün“ bedeutet, es kommt durch Einsatz der approximativ 7500 konserviertesten TFBSn zu einer Erhöhung der Anzahl der vorhandenen Kernsequenz der Bindungsstellen einer Matrize, „gelb“ erbrachte keine Veränderung und „rot“ markiert diejenigen Matrizen, deren dazugehörige TFBSn durch Einsatz der approximativ 7500 TFBSn/Matrize im Vergleich zum auto-score eine Verminderung der Anzahl der TFBSn mit Kernsequenz aufweisen.

Durch Einsatz der konserviertesten approximativen 7500 TFBSn pro Matrize zeigen 21 Matrizen im Vergleich zu der Verwendung des auto-scores ein vermehrtes Auftreten der Kernsequenz in den detektierten TFBSn, 10 Matrizen zeigen keine Veränderung in den angegebenen Grenzen und 11 Matrizen weisen eine Verminderung der Anzahl der TFBSn mit konservierter Kernsequenz auf. Dies lässt die Schlussfolgerung zu, dass durch das

manuelle Einstellen der ausgegebenen TFBSn von ca. 7500 die Anzahl Bindungsstellen einer Matrize mit vorhandener Kernsequenz prinzipiell erhöht wird.

Da die Analysen belegen, dass durch die Verwendung der approximativ 7500 Bindungsstellen pro Matrize in den meisten Fällen eine höhere Anzahl an Bindungsstellen die Kernsequenz erzielt werden kann, wurde bei den folgenden Analysen die Einstellung auf die approximativ 7500 konserviertesten TFBSn bei allen Versuchsgruppen eingesetzt.

2.5 Entwickelte Software

Im Rahmen dieser Arbeit wurden zahlreiche Programme entwickelt, um z.B. die Genomdaten von *Arabidopsis thaliana* zu verarbeiten und um Nukleotide aus dem Genom zu extrahieren oder auch um die Verteilungsanalysen mit den putativen Transkriptionsfaktor-Bindungsstellen durchführen zu können.
Die entwickelten Perl-Skripte wurden bereits erwähnt. Diese sind CGparseTIGR.pl, CGCleanForImport.pl, CGJoinFilesAddID.pl, SelectModels.pl, CGFeatureUpDown.pl und ExtractSequencesChooseReverseFeature.pl.
Des Weiteren wurden Visual Basic Programme entwickelt, auf die im Folgenden eingegangen wird.

2.5.1 SeqToMatrix

SeqToMatrix ist ein Visual Basic Programm, das entwickelt wurde, um aus einzelnen DNA-Sequenzen eine Matrize erstellen zu können. Die Sequenzen, aus denen die Matrize berechnet wird, können entweder direkt durch den Benutzer eingegeben werden oder können, sollten Zufallssequenzen für spätere Analysen benötigt werden, mit dem Programm selbst erzeugt werden. Hierbei können entweder rein zufällige oder Sequenzen, die einem benutzerdefinierten AT-Gehalt entsprechen, generiert werden. Die Anzahl der von SeqToMatrix zu erzeugenden Sequenzen und deren Länge können vorher festgelegt werden.
Eine weitere Funktionalität des Programms beinhaltet die Bestimmung der revers komplementären Sequenz einer einzelnen DNA-Sequenz, wonach diese bei den für die Matrizenerstellung zu verwendenden Sequenzen eingefügt werden kann.
Die Matrize wird erzeugt, indem die Anzahl der Nukleotide an jeder vorhandenen Position aufsummiert wird. Hierbei wird vom Programm kein Alignment erzeugt, das heißt die Sequenzen werden nicht gruppiert, sondern es verarbeitet zeichenweise die vorhandenen vertikalen Positionen der eingegebenen Sequenzen.

Zusätzlich zu der ausgegebenen Matrize wird die Konsensus-Sequenz der Matrize berechnet. Diese Sequenz fasst Mehrdeutigkeitspositionen, das heißt, dass nicht nur ein bestimmtes Nukleotid präferenziell auftritt sondern dass auch andere Nukleotide an der entsprechenden Position vorkommen können, mit einem festgelegten Buchstabencode zusammen. Die Regeln zu der Ermittlung der Konsensus-Sequenz wurden nach Cavener, 1987, implementiert (Kapitel 1.3).
Des Weiteren wird die Anzahl der Nukleotide pro Position aus der Matrize aufsummiert und ausgegeben.
Mit dem Programm ist es auch möglich, den AT- und GC-Gehalt einer vorhandenen Matrize, die z.B. aus der Literatur extrahiert oder mit SeqToMatrix selbst erzeugt wurde, zu bestimmen.

2.5.2 DatabaseViewer

Der DatabaseViewer ist ein Visual Basic Programm, welches dazu entwickelt wurde, die in Kapitel 2.1 erzeugten Daten bezüglich der Positionsangaben der einzelnen Bereiche innerhalb von Genen zu überprüfen. Aus diesen Daten wurde die Datenbanktabelle cds_all erzeugt. Auf die Beschreibung der einzelnen Kontrollen wird hier verzichtet. Die Datenbanktabelle cds_all enthielt zunächst keine Datensätze bezüglich der intergenischen Bereiche. Diese Daten wurden mit dem DatabaseViewer erzeugt und in die Tabelle mit aufgenommen.

2.5.3 TFBSites Datenbank und die TFBSitesTools

Für die Analyse der Verteilung von Transkriptionsfaktor-Bindungsstellen wurde eine Datenbank benötigt, um diese Bindungsstellen zu verwalten und abfragen zu können. Ursprünglich wurde sowohl die Datenbank TFBSites als auch die Visual Basic Software, welche auf die Datenbank zugreift, von Lorenz Bülow (pers. Mitteilung) entwickelt. Im Verlauf dieser Arbeit wurden beide Ressourcen weiter entwickelt und angepasst. Erläutert wird hier einerseits der Aufbau der Datenbank an sich, andererseits jedoch nur die für die Analysen relevanten Funktionalitäten der TFBSitesTools.

Die Datenbank TFBSites befindet sich auf einem Microsoft SQL-Server 2000 an der TU Braunschweig im Institut für Genetik. Die Haupttabellen der TFBSites Datenbank sind hierarchisch aufgebaut. Dies wurde durchgeführt, um die enorme Menge an zu erwarteten Datensätzen übersichtlich zu verwalten und Abfragezeiten zu verkürzen. Die catalog-Tabelle, die aus den Attributen table_no, table_name und creation aufgebaut ist, ist die hierarchisch

höchste Tabelle und beinhaltet die Namen der nächstniedrigeren Hierarchie-Tabellen und dient dazu, einen Überblick über diese Tabellen zu erlangen. Diese Tabellen können benutzerdefiniert benannt werden und enden generell mit „cat“ (z.B. DOFcat) als Hinweis darauf, dass diese Tabellen mit der catalog-Tabelle im Zusammenhang stehen. Die Spaltennamen sind import_no, factor_name, description, number, creation, status, matrix_length, core_sequence und core_pos. Diese Tabellen enthalten generelle Informationen über die Suche nach Transkriptionsfaktor-Bindungsstellen mit einer bestimmten Matrize. Die hierarchisch niedrigste Tabelle enthält den benutzerdefinierten Namen der nächsthöheren Tabelle ohne den Zusatz „cat“ (z.B. DOF). Diese Tabelle enthält die eigentlichen Bindungsstellen und zusätzliche Informationen zu diesen einzelnen Bindungsstellen. Sie beinhaltet die Attribute factor_no, factor_name, chromosome, position, strand, score, lnp, redundant, id, sequence, core, gene_name, gene_position, gene_strand, temp_redundant, tss_position, feature_location, feature_up_down und feature_position.
Des Weiteren umfasst die Datenbank die Tabellen cds und cds_all. Diese Tabellen enthalten die Informationen bezüglich der genomischen Positionen der Gene, die mit dem Perl-Skript CGparseTIGR.pl (Kapitel 2.1) aus dem TIGR Release 5 extrahiert wurden. Die cds Tabelle besteht aus den MODEL-Zeilen der Gene, welche den jeweiligen Translationsstartpunkt und -endpunkt der Gene protokolliert. Diese Informationen wurden mit SQL-Befehlen modifiziert, so dass sie mit dem Format der bestehenden TFBSites übereinstimmen. Die Spaltennamen der cds sind id, type, name, contig, start_pos, end_pos, strand, atg_pos, tss_pos, help_pos, tag, evidence und comment.

Die cds_all-Tabelle stellt die Erweiterung der cds Tabelle dar, denn sie umfasst die Positionsangaben der UTRs, Introns und Exons der Gene. Die Informationen wurden analog zu denen für die cds Tabelle erlangt, und die Daten wurden aufbereitet, so dass sie mit dem Format der TFBSites kompatibel waren, wobei hier noch ergänzende Spalten eingefügt wurden. Die Attribute der cds_all Tabelle sind id, name, chromo, general_type, sub_type, orientation, start_pos, stop_pos, atg_pos, tss_pos, feature_number, up_down, description, evidence und comment.

Gene können mehrere Splice-Varianten aufweisen. Um einen systematischen Fehler bei den Analysen zu vermeiden, wurde nur jeweils nur eine Variante verwendet. Dieses war in allen Fällen die mit der kleinsten Zahl (z.B. bei At1g55490.1 und At1g55490.2 wurde At1g55490.1 verwendet). Die Information darüber, welche Variante verwendet werden soll, wurde bei beiden cds Tabellen in der Spalte description hinterlegt. Generell wurden mit einem SQL-Befehl alle „.1“ Varianten markiert. Einige Gene weisen jedoch keine „.1“ Variante auf. Diese

wurden von Nils Ole Steffens (pers. Mitteilung) identifiziert. In dem Fall wurde die nächsthöhere Splicevariante markiert.

Nachdem nun die Datenbank an sich beschrieben wurde, erfolgt nun die Erläuterung der auf die Datenbank zugreifenden TFBSitesTools, die von einem lokalen Computer aus verwendet werden können.

Die TFBSitesTools stellen einen Visual Basic Client dar, mit dem die mit Patser (Kapitel 2.4) detektierten putativen Transkriptionsfaktor-Bindungsstellen in die Datenbank importiert werden können, aber sie dient auch dazu, die in die Datenbank eingebrachten Daten weiter zu bearbeiten, sie abzufragen und Analysen durchführen zu können. Zusätzlich können mit diesen Tools Zufallspositionen im Genom erzeugt werden. Auf die einzelnen Funktionalitäten wird nun eingegangen.

Erzeugung von Zufallspositionen im Patser Format

Zufallspositionen im *Arabidopsis thaliana* Genom können mittels Aktivierung des Buttons „Create random TFBS“ erzeugt werden. Hierbei ist es möglich, das Vorkommen von identischen Positionen auszuschließen. Die Ausgabe der erzeugten Positionen erfolgt in einer automatisch generierten Textdatei im Patser-Format, so dass diese Positionen anschließend in die Datenbank TFBSites importiert werden können.

Import von Transkriptionsfaktor-Bindungsstellen

Liegen Bindungsstellen und Positionen im Patser-Format in einer Datei vor, können sie mittels „Import“ in die Datenbank importiert werden. Hierbei ist es möglich, neue Tabellen („name“cat) in der Datenbank anzulegen. Der Name des Transkriptionsfaktors bzw. eine Bezeichnung für die Matrizensuche müssen vor dem Import eingegeben werden. Des Weiteren ist es sinnvoll, eine nähere Beschreibung der Parameter, mit denen die Matrizensuche durchgefürt wurde, einzugeben, welche in die Datenbank mit aufgenommen werden kann.

Redundanzanalyse

Es ist möglich, dass Bindungsstellen einer Matrizensuche identische genomische Positionen belegen. In diesen Fällen wurde nur eine in die Analysen mit einbezogen, um eine Überrepräsentierung zu vermeiden. Dies bedeutet, dass sowohl eine putative Transkriptionsfaktor-Bindungsstelle auf dem annotierten wie auch auf dem nicht annotierten DNA-Strang vorliegt, die Kernsequenzen jedoch deckungsgleich sind. Um das Ausmaß von identischen genomischen Positionen zu erfassen, können die TFBSitesTools eingesetzt

werden. Hierzu wird zunächst der Button „Analyse one factor“ aktiviert. Danach kann die zu analysierende Matrizensuche ausgewählt werden (Select factor). Über die Eingabe eines benutzerdefinierten scores kann die Anzahl der ausgewählten Bindungsstellen aus der Datenbank eingeschränkt werden. Wird anschließend bei den angegebenen Tabs „Analyse“ selektiert, kann der Benutzer den erlaubten Abstand von zwei putativen Transkriptionsfaktor-Bindungsstellen (oder generell Positionen) eingeben, um noch in die folgende Analyse mit einzugehen. Danach stehen die Parameter für die Analyse fest, und der Button „Analyse“ kann aktiviert werden, um die Berechnung durchzuführen. Das Ergebnis erscheint in einem entsprechenden Textfeld. Der Abstand zwischen zwei putativen Transkriptionsfaktor-Bindungsstellen wird nukleotidgenau angegeben. Hierbei werden zudem Orientierungen protokolliert.
Das Ergebnis der Analyse kann genutzt werden, um identische putative Transkriptionsfaktor-Bindungsstellen in der Datenbank zu markieren, so dass von zweien nur eine in künftige Analysen mit eingehen kann. Um dieses zu erreichen, kann im Optionsfeld die Orientierung der beiden Bindungsstellen ausgewählt werden, die an identischen genomischen Positionen liegen, wonach mit einer Aktivierung der Schaltfläche „Tag records“ die Berechnung durchgeführt und das Ergebnis in der Datenbank gespeichert wird.

Eine Redundanzanalyse, unter Berücksichtigung der resultierenden Konsensus-Sequenzen von Alignment Matrizen, ist beispielsweise in Steffens *et al.,* 2005, publiziert. Hierbei wird der Abstand von zwei Transkriptionsfaktor-Bindungsstellen einer Matrizensuche ermittelt, bei dem identische genomische Positionen auftreten können. Er tritt genau dann auf, wenn auf dem annotierten sowie auf dem nicht annotierten DNA-Strang die Kernsequenzen, ermittelt aus der Konsensus-Sequenz, deckungsgleich sind. Vor allem tritt dies bei palindromischen Sequenzen auf.
In dieser Arbeit wurde die Redundanzanalyse analog zu Steffens *et al.,* 2005, durchgeführt. Die Bearbeitung der Konsensus-Sequenzen (Erzeugung der revers komplementären Sequenz) erfolgte mit SeqToMatrix (Kapitel 2.5.1). Mit den TFBSitesTools wurde von zwei identischen genomischen Positionen nur eine in die folgenden Analysen mit einbezogen.

<u>Einlesen und Analyse der Kernsequenz</u>
Generell ist für die Fähigkeit von Transkriptionsfaktoren, an die entsprechende DNA-Sequenz binden zu können, die Konservierung der Kernsequenz entscheidender als die Konservierung der flankierenden Nukleotide. Mit den TFBSitesTools ist es möglich, die importierten putativen Transkriptionsfaktor-Bindungsstellen auf das Vorhandensein der Kernsequenz hin zu analysieren. Voraussetzung dafür ist, dass die chromosomalen DNA-Sequenzen lokal auf dem Rechner vorliegen.

Um die Kernsequenzen zu bestimmen, wird zunächst der Button „Analyse one factor“ betätigt, wonach in dem Tab „Select“ die entsprechenden Datensätze ausgewählt werden können. Im Tab „Core region/CDS“ erfolgt dann die Festlegung der Länge der Matrize, die Kernsequenz (z.B. ACGT) sowie die Position des ersten Kernsequenznukleotids vom Beginn der Sequenz aus betrachtet. Die Aktivierung des Buttons „Core region“ startet die Berechnung. Die generellen Informationen über die Kernsequenz werden in der entsprechenden cat-Tabelle in der Datenbank (Spalten matrix_length, core_sequence, core_pos) gespeichert. Für jede einzelne Bindungsstelle der Matrizensuche erfolgt das Einlesen der DNA-Sequenz, die mit der Matrize identifiziert wurde (Spalte sequence), sowie die Feststellung, ob in dieser Sequenz die Kernsequenz auftritt, was durch eine Markierung in Form eines festgelegten Wertes in einer Spalte der Tabelle hinterlegt wird.

Nach dem Einlesen der DNA-Sequenz der einzelnen Bindungsstellen und der Markierung von Bindungsstellen, die die Kernsequenz aufweisen, kann über „Select factor“ durch Aktivieren bzw. Deaktivieren der Option „exclude non-core sequence“ festgestellt werden, wie viele putative Transkriptionsfaktor-Bindungsstellen der Datensätze die Kernsequenz aufweisen.

<u>Einlesen des nächstgelegenen Gens einer putativen Transkriptionfaktor-Bindungsstelle sowie Analyse der Bindungsstellen relativ zum nächstgelegenen Gen</u>

Nach dem Import von putativen Transkriptionsfaktor-Bindungstellen (oder generell Positionen) können die dazugehörigen nächstgelegenen Gene eingelesen werden. Das nächstgelegene Gen ist hier genau das, dessen Translationsstartpunkt der putativen Transkriptionsfaktor-Bindungsstelle am nächsten liegt. Die Funktionalität kann genutzt werden, indem die entsprechende Matrizensuche über den Button „Analyse one factor im Tab „Select factor“ zunächst ausgewählt wird, wonach im Tab „Core region/CDS“ der Button „ATG/TSS“ betätigt wird. Im Verlauf der Berechnung werden der Genname des nächstgelegenen Gens (gene_name), der DNA-Strang (strand), auf dem dieses Gen annotiert ist, sowie der Abstand der putativen Transkriptionsfaktor-Bindungsstelle relativ zu dem Translationsstart (atg_position) in der Datenbank protokolliert. Zusätzlich wird, wenn dieses Gen einen annotierten Transkriptionsstart aufweist, die relative Distanz dieser Bindungsstelle zum Transkriptionsstartpunkt (tss_position) in die Datenbank mit aufgenommen.

Ist diese Vorkalkulation durchgeführt, kann die Analyse von putativen Transkriptionsfaktor-Bindungsstellen relativ zum nächstgelegenen Gen durchgeführt werden. Zu diesem Zweck werden ausgehend vom Button „Analyse one factor“ die Bedingungen für die Analyse

gewählt. Hierzu zählt die Matrizensuche an sich, wobei über die benutzerdefinierte Eingabe des score die Anzahl der putativen Transkriptionsfaktor-Bindungsstellen eingeschränkt werden kann. Ausgehend von einer vorherigen Redundanzanalyse ist die automatische Auswahl der Checkbox „exclude redundant records“ zu belassen. Sollte mit diesen Datensätzen jedoch eine Kernsequenzanalyse durchgeführt worden sein, sollte die Checkbox „exclude non-core sequence“ deaktiviert werden. Im Tab „Analyse (ext.)“ können dann die Transkriptionsfaktor-Bindungsstellen relativ zum nächstgelegenen Gen mit „Count (ATG)“ untersucht werden, wobei als Bezugspunkt der Translationsstart des nächstgelegenen Gens verwendet wird. Alternativ gehen mit „Count (TSS)“ nur die Transkriptionsfaktor-Bindungsstellen in die Analyse mit ein, deren dazugehöriges Gen einen annotierten Transkriptionsstartpunkt aufweist. Das Ergebnis wird in Form einer Texttabelle ausgegeben. Einerseits werden hierbei für jede Distanz die Anzahl der putativen Transkriptionsfaktor-Bindungsstellen angegeben, andererseits werden diese Distanzen auch in 50 Nukleotid große Bereiche aufgeteilt und die einzelnen putativen Bindungsstellen entsprechend aufsummiert.

Einlesen der Lokalisierung der putativen Transkriptionsfaktor-Bindungsstellen relativ zu den einzelnen Bereichen der Gene sowie Analyse der Bindungsstellen relativ zu den einzelnen Bereichen der Gene

Um Verteilungsanalysen von Transkriptionsfaktor-Bindungsstellen relativ zu bestimmten Bereichen innerhalb von Genen zu ermöglichen, ist es sinnvoll, diese Daten vorzukalkulieren. Hierzu können die TFBSitesTools eingesetzt werden. Hierbei werden für jede putative Transkriptionsfaktor-Bindungsstelle einer Matrizensuche die Bereiche innerhalb von Genen protokolliert (Intron, Exon, 5’UTR, 3’UTR), sowie die entsprechende Nummer des Bereichs innerhalb des jeweiligen Gens (Kapitel 2.1.1) und die relative Distanz zum Beginn dieses Bereiches (Spalten feature_location, feature_up_down und feature_position) mit in die jeweilige Tabelle aufgenommen. Als Voraussetzung für die Vorkalkulation ist das Einlesen jeder Bindungsstelle relativ zu ihrem nächstgelegenen Gen notwendig, siehe vorheriger Abschnitt.

Für das Einlesen der Lokalisierung jeder putativen Transkriptionsfaktor-Bindungsstelle (oder generell Position) relativ zu den einzelnen Bereichen innerhalb der Gene ist zunächst die Auswahl der Matrizen, die über „Analyse one factor“ im Tab „Select factor“ notwendig. Anschließend kann im Tab „Core region/CDS“ der Button „Features“ gewählt werden, der die Vorkalkulation auslöst. In die Datenbank werden die relativen Distanzen bis zum Beginn des transkribierten oder tranlatierten Bereichs des zuvor zugeordneten Gens ermittelt. Dabei ist

es unerheblich, ob sich eine Bindungsstelle innerhalb eines Gens oder im intergenischen Bereich befindet.

Ist das Einlesen beendet, können die Parameter der Verteilungssanalyse im Tab „Select factor“ eingestellt werden (score, Checkboxen bezüglich der Redundanz und der Kernsequenz). Im Tab „Analyse (ext.)“ ist die Auswahl des zu analysierenden Bereichs über Auswahlfelder möglich. Wahlweise können alle Bereiche eines Typs im intergenischen Bereich oder innerhalb der Gene (Introns, Exons, 3’UTR, 5’UTR) analysiert werden oder nur ganz bestimmte Bereiche innerhalb von Genen (1. Intron, usw.).

Geht es um eine allgemeine Analyse von Bindungsstellen in einzelnen Bereichen, sollen also alle annotierten TFBSn, die sich z.B. im ersten Intron von Genen befinden, mit einbezogen werden, ist für solch eine Analyse der Betrachtungsbereich „Gene output“ leer zu lassen. Die Analyse selbst wird durch Aktivierung des Buttons „Count“ ausgelöst. Die Ausgabe erfolgt in einer automatisch generierten Textdatei, in der zu jeder einzelnen Distanz, angegeben in Nukleotiden, die Anzahl der vorgefundenen annotierten Transkriptionsfaktor-Bindungsstellen angegeben wird. Zusätzlich dazu findet automatisch auch eine Einteilung der Distanzen in 50 bp-Schritten statt, in denen ebenfalls die vorgefundenen putativen TFBSn aufsummiert werden.

Weiterhin können auch Bindungsstellen, die in ganz speziellen Distanzen relativ zum Start des gewählten Bereichs vorkommen, näher analysiert werden. In dem Fall können die Textfelder bei „Gene output“ mit den Distanzgrenzen, die eine putative TFBS relativ zu dem bestimmten Bereich aufweisen muss, um in die Analyse mit einzugehen, gefüllt werden. Wird anstelle von Distanzgrenzen nur ein Wert angegeben, werden nur Bindungsstellen mit der exakten eingegebenen Distanz untersucht. Wird der „Gene output“ hingegen ausgefüllt und die Analyse mit „Count“ gestartet, werden weitere Analysemöglichkeiten durch Eingabefelder abgefragt. Die erzeugte Ausgabedatei enthält dann neben jeder einzelnen, den Parametern entsprechenden, Bindungsstelle die relative Positionsangaben zu dem Gen, in dessen Bereich die Bindungsstelle lokalisiert ist bzw. welches der Bindungsstelle im Fall des intergenischen Bereichs am nächsten ist. Des Weiteren werden die Sequenz der Bindungsstelle und wahlweise die um einen benutzerdefinierten Bereich, beidseitig erweiterte DNA-Sequenz, angegeben in Nukleotiden, ausgegeben. Diese Sequenz kann eingesetzt werden, um mit anderen Programmen weitere Analysen durchzuführen, mit dem Ziel, weitere Motive zu finden, um eine mögliche Kolokalisation von weiteren Transkriptionsfaktor-Bindungsstellen feststellen zu können.

Einfache Lokalisierung von Transkriptionsfaktor-Bindungsstellen in den bestimmten Bereichen innerhalb der Gene und im intergenischen Bereich

Mit den TFBSitesTools ist es möglich, eine vereinfachte Analyse bezüglich der Lokalisierung von Transkriptionsfaktor-Bindungsstellen in den bestimmten Bereichen von Genen und auch im intergenischen Bereich durchzuführen. Hierbei wird nur festgestellt, in welchem Bereich die Bindungsstelle (oder prinzipiell Position) liegt. Die Voraussetzung für die Nutzung dieser Funktionalität ist das Einlesen der nächstgelegenen Gene der putativen Bindungsstellen sowie die Annotation der Distanz der Bindungsstellen bezüglich der einzelnen Bereiche der Gene. Sind die Voraussetzungen erfüllt, kann zunächst der Button „Analyse one factor“ betätigt werden. Danach kann dann die zu analysierende Matrizensuche ausgewählt werden, incl. der für die Analyse festzulegenden Parameter wie score usw. Anschließend kann der Button „Count (TFBS)“ im Tab „Analyse (ext.)“ verwendet werden. Die Anzahl der Bindungsstellen in den einzelnen Bereichen wird jeweils aufsummiert. Die Ausgabe erfolgt in einer automatisch generierten Textdatei.

Extraktion der höchsten Bereiche der Gene

Für die Analyse der Nukleotidkomposition der einzelnen Bereiche der Gene wurden per SQL-Abfragen im Microsoft Query Analyser die entsprechenden Zeilen in der cds_all Tabelle abgefragt. Die einzelnen Bereiche, z. B. Introns, eines Gens liegen durchnummeriert in der Datenbank vor. Für die Extraktion der am höchsten nummerierten Bereiche wurden die TFBSitesTools eingesetzt.

Für die Feststellung der höchsten Bereiche der Gene ist der Button „Extract“ einsetzbar, erreichbar unter „Analyse one factor“, wonach der Tab „Analyse (ext.)“ ausgewählt werden muss. Über dem „Extract“ Button kann der grundsätzliche Bereich der Gene (Intron, Exon, 5’UTR, 3’UTR) ausgewählt werden, für die das jeweils höchste ermittelt werden soll. Die Ausgabe erfolgt in Form einer automatisch generierten Textdatei, welche analog zu den SQL-Abfragen bezüglich der vorgegebenen Nummer eines Bereichs die entsprechenden Datensätze aus der cds_all Tabelle enthält. Die weitere Verwendung der so generierten Textdatei ist in Kapitel 2.1.2 beschrieben.

Ermittlung der Nukleotidanzahl innerhalb von Genbereichen und im intergenischen Bereich

Für eine generelle Analyse der Anzahl der Nukleotide in den einzelnen Bereichen der Gene sowie im intergenischen Bereich können die TFBSitesTools eingesetzt werden. Zu diesem Zweck kann, ausgehend von „Analyse one factor“, im Tab „Analyse (ext.)“ der Button „Count (Genes)“ verwendet werden. Die Verteilung der Nukleotide wird in einer Textdatei protokolliert und enthält zusätzlich für alle auftretenden Bereiche innerhalb von Genen die Anzahl der Gene, die einen bestimmten Bereich, wie z.B. ein erstes Intron, aufweisen.

2.5.4 BuildGeneQuery

Das Visual Basic Programm BuildGeneQuery wurde entwickelt, um ausgehend von den mittels der TFBSitesTools extrahierten Gennamen (Kapitel 2.5.3) die Funktion dieser Gene in der Datenbank CGAthaMap herauszusuchen. Diese Datenbank ist die Grundlage für die AthaMap (Kapitel 1.5).

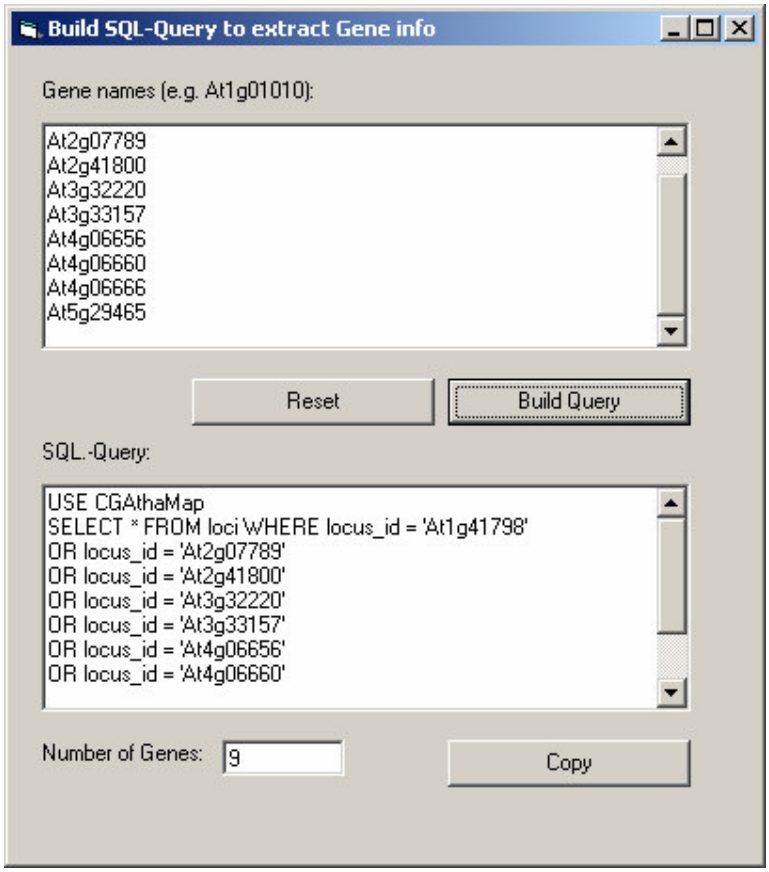


Abb. 9: Screenshot der Benutzeroberfläche des Visual Basic Programms BuildGeneQuery. Die generierte SQL-Abfrage kann direkt im MS Query Analyser verwendet werden.

Die Genliste wurde dabei zunächst in das Programm Notepad2 überführt, wo mittels suchen/ersetzen ab dem Punkt im Gennamen überflüssige Zeichen entfernt wurden. Anschließend erfolgte die Eingabe der bearbeiteten Liste in das Programm BuildGeneQuery, welches die SQL-Query für die Abfrage der Beschreibung der Gene automatisch bildet (Abb. 9). Mit dem Microsoft Query Analyser wurden die Genfunktionen abgefragt.

2.6 Statistik

Im Zusammenhang mit den Verteilungsanalysen wurden verschiedene statistische Formeln, Tests und Programme eingesetzt, um die Qualität der Ergebnisse einschätzen zu können. Im Folgenden wird im Einzelnen auf die verwendete Statistik eingegangen.

2.6.1 Mittelwert, Standardabweichung und Fehler des Mittelwertes

Zur Berechnung des Mittelwertes einer Stichprobe wurde Formel 3 eingesetzt (Bartsch 1994).

Formel 3: Arithmetisches Mittel

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

mit

$\bar{x} = \text{Mittelwert}$

$n = \text{Stichprobenumfang}$

$x_i = \text{Einzelwert}$

Die empirische Standardabweichung σ als Fehler eines Einzelwertes vom Mittelwert ist in Formel 4 definiert.

Formel 4: Standardabweichung

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

Aus der Standardabweichung sowie dem Probenumfang lässt sich der Fehler des Mittelwertes berechnen (Formel 5).

Formel 5: Fehler des Mittelwertes

$$\sigma_m = \frac{\sigma}{\sqrt{n}}$$

2.6.2 Test auf Normalverteilung und gleiche Verteilung mit STATeasy

STATeasy ist ein Programm, mit dem statistische Auswertungen durchgeführt werden können. Im Rahmen dieser Arbeit wurde es verwendet, um eine Prüfung auf

Normalverteilung sowie um Vergleiche zweier Verteilungen miteinander durchzuführen. Es basiert auf der dritten Ausgabe des Lehrbuchs von Lozán und Kausch (2004).

Bei den Verteilungsanalysen relativ zum nächstgelegenen Translationsstartpunkt wurde mit 42 x 7500 Zufallspositionen gearbeitet, deren jeweilige relative Distanz zum Bezugspunkt in 50 bp Schritten zusammengefasst und grafisch dargestellt wurde. Um nun zu überprüfen, ob eine augenscheinliche Normalverteilung der Werte tatsächlich vorliegt, wurde STATeasy eingesetzt. Für die Überprüfung der Normalverteilung wurde der Kolmogoroff-Smirnoff Anpassungstest (modifiziert nach Kuiper) eingesetzt (Lozán und Kausch 2004, Seite 77ff). Bei diesem Test werden der größte und der kleinste Abstand zweier Kurven miteinander verglichen, wobei die Summe einen kritischen Wert nicht überschreiten darf. Beim Vergleich werden die erwarteten Werte bei einer Normalverteilung mit der empirischen Beobachtung eingesetzt.
Bei einem weiteren Test auf Normalverteilung wurde der Schnelltest nach David und Mitarbeitern verwendet (Lozán und Kausch 2004, Seite 79ff). Diese verwendet einen Quotienten untersucht und stellt kritische Schranken auf, außerhalb derer eine Normalverteilung abzulehnen ist.

Zu STATeasy gehört eine Excel-Tabellenkalkulationsdatei, die bei Start des Programms automatisch geöffnet wird. In diese Datei müssen die zu untersuchenden Werte eingefügt werden. Für die Durchführung der Tests wurde, ausgehend von der Startseite von STATeasy, das Modul „STAT11-Einstichprobe“ gewählt, wonach die Option „Prüfung auf Nicht-Normalverteilung“ angeklickt und der Button „Weiter/Eingabe der Daten“ aktiviert wurde. Anschließend wurde im Menü die Rubrik „Datentyp“ und dort „Wertepaare“ ausgewählt. In der Excel-Datei können nun die beiden Spalten der Wertepaare markiert werden, wonach mit „OK“ im Eingabeformular die Eingabe bestätigt werden muss. Anschließend erfolgt die Eingabe einer Namensbezeichung für die Wertepaare, die es dann zu bestätigen gilt. Es folgt die Ausgabe eines Protokolls der Eingabe durch STATeasy, die bestätigt werden muss. Nun werden die für Wertepaare in Frage kommenden statistischen Tests auf Normalverteilung überprüft. Es wurden hier einmal der Kolmogoroff-Smirnoff Anpassungstest (modifiziert nach Kuiper) sowie der Test nach David *et al.* ausgewählt. Die danach abgefragte Irrtumswahrscheinlichkeit betrug in beiden Fällen 5 %, was der Standardeinstellung entspricht. Das Ergebnis des jeweiligen Tests wird anschließend angezeigt und kann exportiert werden.

Neben der Prüfung auf Normalverteilung wurde bei den Zufallspositionen ein Vergleich zweier Verteilungen unter der Voraussetzung, dass die Stichproben unabhängig voneinander

sind, durchgeführt. Verfahren wurde dabei analog zum Schema in Lozán und Kausch 2004, Seite 91. Verwendet wurden hierbei der U-Test, der einen Vergleich der Mittelwerte durchführt (Seite 103ff), sowie der Median-Test, der einen Vergleich der Medianwerte umfasst (Seite 110f).
Zu diesem Zweck wurde zunächst „STAT21- Vergleich zweier Stichproben“ ausgewählt und danach „Daten/Eingabe“ aktiviert. Anschießend erfolgte im Menü die Festlegung auf Datentyp „Einzelne Werte“, wonach der Button „Weiter/Eingabe der Daten“ betätigt und die zu analysierenden Werte in den zwei Spalten separat markiert und mit einem Namen versehen wurden. Das nun ausgegebene Protokoll von STATeasy wurde bestätigt. Anschließend wurde durch Aktivieren des Buttons „Wahl des TESTS“ der U-Test mit der Alternativhypothese „zweiseitiger Test“ gestartet. Hierbei wird zunächst die Voraussetzung überprüft, ob die Daten nicht normalverteilt sind (Kolmogoroff-Smirnoff Anpassungstest, Irrtumswahrscheinlichkeit = 5 %). Es erfolgt die Ausgabe des Ergebnisses dieses Tests, welche mit „Weiter“ bestätigt wurde. Danach wurde im Menü der Button „Ergebnisse“ aktiviert, wonach eine Irrtumswahrscheinlichkeit von 5 % für den U-Test angesetzt wurde. Hiernach wird das Ergebnis des U-Tests angezeigt. Ausgehend von den Ergebnissen wurde nachfolgend der Median-Test mit der Alternativhypothese „zweiseitiger Test“ und einer Irrtumswahrscheinlichkeit von 5 % ausgewählt, wonach das Ergebnis der Berechnungen angezeigt wurde.

2.6.3 P-Wert zur Ermittlung von Signifikanzen

Bei den Verteilungsanalysen wurde das Auftreten von putativen Transkriptionsfaktor-Bindungsstellen (TFBSn) bei diskreten Distanzen relativ zum Start von bestimmten Genbereichen untersucht. Hierbei ist es jedoch notwendig, eine statistische Einschätzung über die Güte der Information zu erhalten. Um zu einem statistischen Kriterium zu gelangen, sind folgende Dinge zu berücksichtigen.

Zunächst sind zwei Hypothesen aufzustellen, die als H_o und H_a bezeichnet werden, wobei sich beide Hypothesen gegenseitig ausschließen.

Formel 6: Nullhypothese

$$H_0 = Nullhypothese = \text{Das Auftreten dieser Anzahl an TFBSn an der Position beruht auf dem Zufall}$$

Formel 7: Alternativhypothese

$$H_a = Alternativhypothese = \text{Das Auftreten der TFBSn an der Position beruht auf einem realen Effekt}$$

Bei jeder einzelnen Distanz können zwei verschiedene Fälle eintreten, entweder TFBSn treten auf oder TFBSn treten eben nicht auf. Es existiert also eine Binominalwahrscheinlichkeit und es kann von einer Binomalverteilung ausgegangen werden. Die Binominalverteilung gilt für Stichproben vom Umfang n mit Zurücklegen, jedoch ist bei großem n das Zurücklegen kein Kriterium und es lässt sich die Bernoulli-Gleichung einsetzen (Bartsch 1994).

Formel 8: Bernoulli-Gleichung

$$P = \binom{n}{k} * p^k * q^{n-k}$$

mit

n = Anzahl der Elemente (hier: Anzahl Nukleotide bei spez. Distanz)

k = Gibt an, aus wie vielen Elementen jede Kombination bestehen soll (hier: Anzahl TFBSn bei spez. Distanz)

p = Wahrscheinlichkeit für das Auftreten des Ereignisses (hier: TFBS tritt bei spez. Distanz auf: p = $Gesamt_{TFBS\ der\ Matrizensuche}$/$Gesamt_{Chromosomenlänge}$)

q = Gegenwahrscheinlichkeit (hier: TFBS tritt bei spez. Distanz nicht auf)

mit

Formel 9: Gegenwahrscheinlichkeit für p

$$q = 1 - p$$

Für die Berechnung von n (Formel 8) an jeder diskreten Distanz wurde das Perl-Script ExtractSequencesChooseReverseFeature.pl (Kapitel 2.1.2) eingesetzt. Das bedeutet, die Anzahl der Nukleotide im Genom, die genau diese Distanz zu einem bestimmten Bereich aufweisen, wurde erfasst.

Der erste Teil von Formel 8 wird nun gesondert betrachtet.

Formel 10: Kombinationen (gelesen: n über k)

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Anmerkung: In Excel ist die Formel zur Berechnung der Kombinationen verfügbar „=KOMBINATIONEN(n;k)“.

Somit sind alle Variablen der Bernoulli-Gleichung identifiziert. Mit dieser Gleichung jedoch würde nur die Wahrscheinlichkeit für das Auftreten von einer diskreten Anzahl von TFBSn berechnet werden können. Im Zusammenhang der Analyse ist jedoch die Wahrscheinlichkeit für das Auftreten von der beobachteten Anzahl oder mehr TFBSn wichtig, so dass die Gleichung um Reihenglieder ergänzt werden muss (Formel 11).

Formel 11: Ergänzte Bernoulli-Gleichung

$$P = \binom{n}{k} * p^{k} * q^{n-k} + \binom{n}{k+1} * p^{k+1} * q^{n-(k+1)} + \binom{n}{k+2} * p^{k+2} * q^{n-(k+2)} + \ldots + \binom{n}{k_{\max}} * p^{k_{\max}} * q^{n-k_{\max}}$$

Da bei dieser Gleichung die Inkremente ab k = 4 kaum noch Auswirkungen auf P haben, wurde bei Einsatz dieses statistischen Kriteriums die Gleichung bis zu k = 4 berechnet.

Es gilt:

Formel 12: Nullhypothese ist gleich der Wahrscheinlichkeit P aus der Bernoulli-Gleichung

$$H_0 = P$$

Nachdem nun der P-Wert bzw. H_o bekannt ist, wird noch ein Maß benötigt, das eine Entscheidung zulässt, ob die Nullhypothese zutrifft oder sie abzulehnen gilt. Diese Entscheidung lässt sich mit dem Signifikanzniveau α treffen, welches die Irrtumswahrscheinlichkeit dafür ist, H_0 zu verwerfen, obwohl die Hypothese zutrifft.

In Analogie zum üblichen Signifikanzniveau in der Biologie wurde für diese Analysen

Formel 13: Festlegung des Signifikanzniveaus

$$\alpha = 0{,}05$$

gesetzt. Das bedeutet, es besteht eine 95 %ige Sicherheit, dass die Entscheidung darüber, die Hypothese zu bestätigen oder zu verwerfen, tatsächlich die richtige Entscheidung war.

Für die Verwendung von Formel 11 bedeutet dies:

Formel 14: Ablehnung der Nullhypothese

$$P < 0{,}05 \Rightarrow P \neq H_0$$

In dem Fall beruht das Auftreten von dieser Anzahl an TFBSn an der diskreten Distanz mit einer Wahrscheinlichkeit von 95 % und mehr nicht auf dem Zufall, sondern ist ein Resultat eines realen Effekts.

2.8 Sequenzdarstellung mit WebLogo

DNA-Sequenzen können mit dem WebLogo grafisch dargestellt werden (Schneider und Stephens 1990). Die ausgegebene Grafik beinhaltet die generelle Konsensussequenz der eingegebenen Sequenzen, wobei das am häufigsten vorkommende Nukleotid an oberster Position aufgeführt ist. Des Weiteren gehen die Frequenzen der Nukleotide und die Qualität der Information an jeder Position der Sequenz, dargestellt in bits auf der Ordinate, in das berechnete Logo mit ein. Maximal werden bei der Eingabe von DNA bzw. RNA-Sequenzen an einer Position zwei bits angegeben. Die Höhe der bits eines Nukleotids wird aus dem Produkt aus der Frequenz des Nukleotids an der entsprechenden Position und dem Informationsgehalt berechnet. Der Informationsgehalt setzt sich aus Formel 15 zusammen (Crooks *et al.,* 2004).

Formel 15: Berechnung des Informationsgehaltes R_{Seq} mit WebLogo

$$R_{Seq} = S_{\max} - S_{obs} = \log_2 N - \left(\sum_{n=1}^{N} p_n \log_2 p_n \right)$$

mit

$S_{\max} = \max imal \;\; mögliche \;\; Entropie$

$S_{obs} = auftretende \;\; Entropie$

$N = Anzahl \;\; Symbole \, (DNA : N \in [a, c, g, t])$

$p_n = Frequenz \;\; n \;\; an \;\; Position$

Bei einer Eingabe von Nukleotiden ist laut Formel 16:

Formel 16: Maximum laut Formel 16 für Nukleotide

$$\log_2 N \;\; (mit \;\; N = 4) = 2$$

was das Maximum von 2 bits auf der Ordinate erklärt.

Die Entropie in Formel 15 ist ein Maß für die Unordnung in einem System. Im weitesten Sinn könnte man die Entropie als Maß für den Zufall betrachten. Daher kann durch den dargestellten Informationsgehalt ein Eindruck über die Zufälligkeit der Nukleotidverteilung an einer Position erhalten werden. Liegt dieser für ein Nukleotid bei beim Maximalwert, im Fall von eingegebenen Nukleotiden bei dem Wert 2, deutet das auf einen realen Effekt hin. Ist der Informationsgehalt gering, sind entsprechend niedrige bit-Werte dargestellt.

Die Sequenzlogos in dieser Arbeit wurden mittels der online Ressource http://weblogo.berkeley.edu/ erzeugt.

2.9 Genexpressionsanalysen mit Genvestigator

Genvestigator ist eine online verfügbare Datenbank, die Microarray-basierte Genexpressionsprofile von *Arabidopsis thaliana* Genen enthält (Zimmermann *et al.,* 2005), und über https://www.genevestigator.ethz.ch/ erreichbar ist. Eingesetzt wurde die Funktionalität „Meta Profile Analysis“ (Abb. 10).

Ausgewählt wurde der Organismus *Arabidopsis thaliana*, als „Array type“ das „ATH1 22k array“ von Affymetrix, wobei alle Array Quellen für die Analysen ausgewählt wurden, jedoch wurden nur Arrays berücksichtigt, deren Qualität hoch war (Abb. 10). Nachdem die Auswahl einen benutzerdefinierten Namen erhalten hatte, wurden die Einstellungen durch „OK“ bestätigt. Anschließend konnten die Gennamen durch Aktivierung des grünen „+“ Buttons (roter Pfeil in Abb. 10) eingegeben werden. Es erfolgte danach eine interne Validierung bezüglich des Vorhandenseins der Gennamen in der Datenbank. Da standardmäßig jedoch bei der Expressionsanalyse durch Genevestigator „Probeset IDs“ statt der Gennamen angegeben werden, wurde durch Aktivierung des mit dem blauen Pfeil markierten Buttons eine Eingabemaske aufgerufen, mittels derer die Ausgabe von Gennamen erzielt werden konnte.

Die Expressionsanalyse wird daraufhin automatisch angezeigt. Benötigt wurden für die hier vorgestellten Analysen „Anatomy“, „Development“ und „Stimulus“. Die angezeigten Grafiken konnten noch angepasst werden. Durch Aktivierung der einzelnen drei Expressionsanalysen im Reiter im oberen Teil von Abb. 10 konnten die Ergebnisse als „heat map“ angezeigt werden.

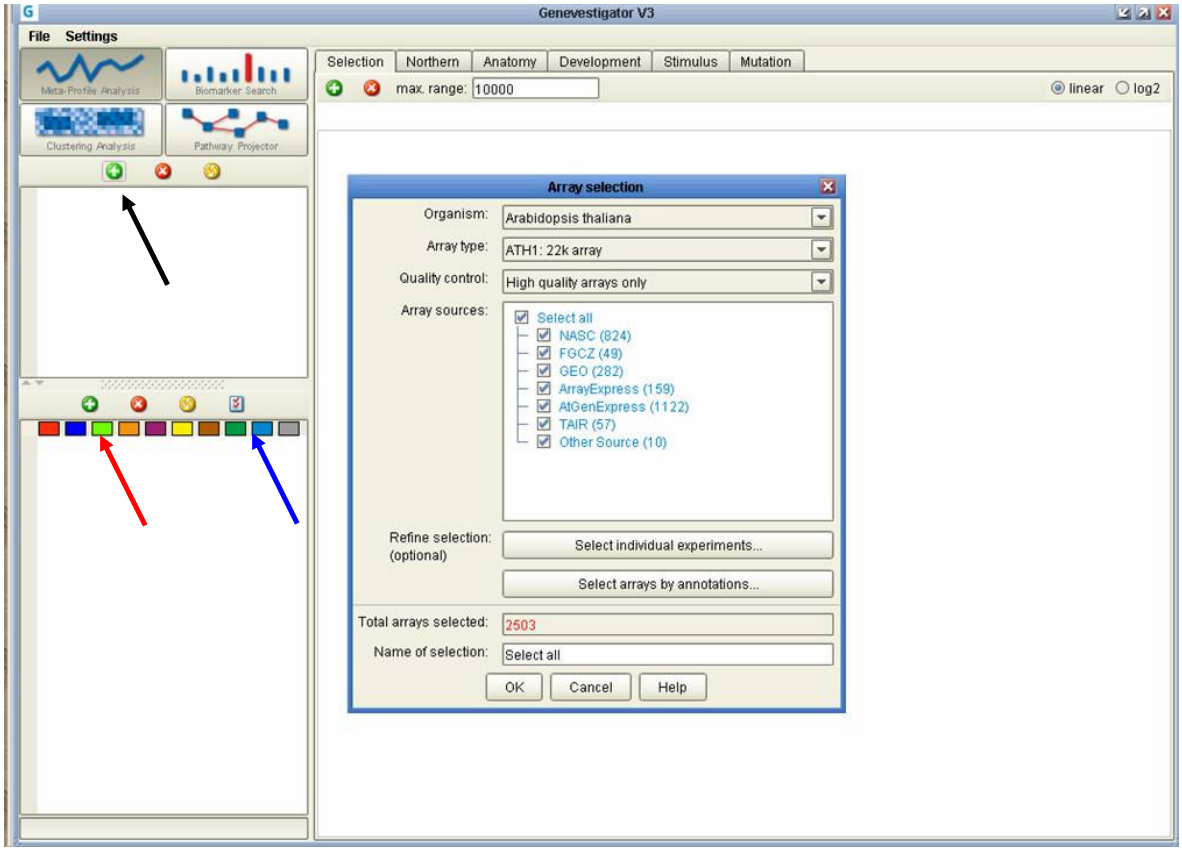


Abb. 10: Screenshot der Startseite bei den Genevestigator Expressionsanalysen, nach Aktivierung des grünen „+“, welches durch den schwarzen Pfeil hervorgehoben wurde, wird das Popup-Fenster „Array selection“ ausgegeben. Die restlichen Pfeile werden im Text erläutert.

Das bedeutet, dass bei der Anatomie und der Entwicklung die Genexpression einen blau-weiß Farbcode hat, wobei blau eine Detektion des Expressionssignals des betreffenden Gens bedeutet und weiß darauf hindeutet, dass kein Signal gemessen werden konnte. Beim Stimulus wurde ein grün-schwarz-rot Farbcode eingesetzt. Die grüne Farbe signalisiert eine Downregulation des entsprechenden Gens, die rote Farbe hingegeben belegt eine Upregulation des Gens und schwarz weist auf ein indifferentes Signal hin.
Im Falle des Stimulus wurde die Anzeige als „tree“ zusätzlich ausgewählt, weil dadurch eine bessere Übersicht gewährleistet wurde.
Die mittels Genevestigator erhaltenen Expressionsgrafiken konnten mit der Export-Funktion lokal gespeichert werden.

3. Ergebnisse

3.1 Analyse der Nukleotidkomposition

Der AT-Gehalt (respektive GC-Gehalt) der DNA-Sequenzen einer Spezies gilt als taxonomisches Merkmal und kann einen Einfluss auf die Verteilung von Transkriptionsfaktor-Bindungsstellen (TFBSn) haben, die ihrerseits ebenfalls einen bestimmten AT-Gehalt aufweisen. Daher wurden im Folgenden die Nukleotidkompositionen im intergenischen Bereich, in der 5'UTR, den Introns, den Exons und der 3'UTR analysiert. Die Untersuchungen wurden auf die erste 5'UTR, die ersten fünf Introns, die ersten vier Exons und die erste 3'UTR ausgeweitet, da Verteilungsanalysen von Transkriptionsfaktor-Bindungsstellen auch in diesen bestimmten Bereichen durchgeführt wurden. Die gesamte 5'UTR beginnt am ersten Nukleotid der 5'UTR und endet beim letzen Nukleotid der 5'UTR. In einer 5'UTR können Introns enthalten sein. Die erste 5'UTR beginnt am ersten Nukleotid der 5'UTR und endet beim letzten Nukleotid vor dem in der 5'UTR enthaltenen Intron. Wenn sie keine Introns enthält, entspricht das letzte Nukleotid der ersten 5'UTR dem der gesamten 5'UTR. Gleiches gilt für die gesamte 3'UTR bzw. die erste 3'UTR. Um die Nukleotidkomposition analysieren zu können, wurden die Start- und Stoppangaben der entsprechenden Bereiche aus der cds_all Tabelle (Kapitel 2.5.3) mit SQL-Abfragen extrahiert, wonach die Nukleotidkomposition mit dem Perl-Skript ExtractSequencesChooseReverseFeature.pl (Kapitel 2.1.2) berechnet und mit Excel grafisch ausgewertet wurde. 16319 Gene besitzen eine 5'UTR1, die Anzahl der Gene mit einem ersten Intron beträgt 21086, mit einem zweiten 17100, mit einem dritten 13998, mit einem vierten 11481 und einem fünften 9346. 29993 Gene weisen ein erstes Exon auf (was zu erwarten ist), 20556 Gene ein zweites, 16673 Gene ein drittes und 13596 Gene ein viertes Exon. Die Anzahl der Gene mit einer 3'UTR1 beträgt 17580.

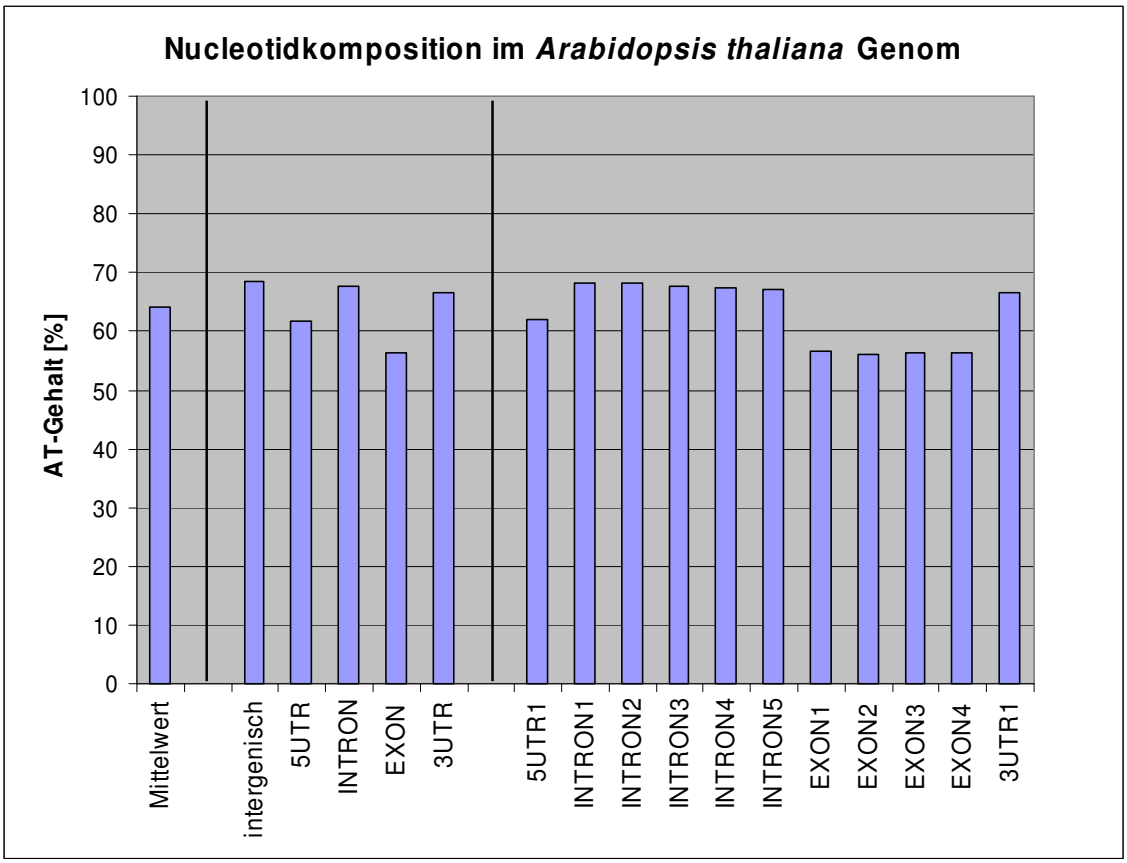


Abb. 11: AT-Gehalt im *Arabidopsis thaliana* Genom. Links ist der durchschnittliche AT-Gehalt im gesamten Genom aufgeführt (Mittelwert). Weiterhin sind die durchschnittlichen Werte für den intergenischen Bereich, für die gesamte 5'UTR, die gesamten Introns, die gesamten Exons und für die gesamte 3'UTR dargestellt. Im rechten Teil der Abbildung sind die Durchschnittswerte für die erste 5'UTR, die ersten fünf Introns, die ersten vier Exons und die erste 3'UTR aufgetragen. Die genauen Werte befinden sich im Anhang in Tabelle 37.

Wie in Abb. 11 dargestellt, beträgt der durchschnittliche AT-Gehalt des *Arabidopsis thaliana* Genoms 64,13 %. Bei der Nukleotidzusammensetzung im intergenischen Bereich, in der 5'UTR, im Intron, im Exon sowie der 3'UTR lässt sich feststellen, dass der AT-Gehalt im intergenischen Bereich am höchsten ist (68,67 %). In Introns (67,59 %) und der 3'UTR (66,47 %) ist der AT-Gehalt höher als in der 5'UTR (61,84 %), wobei er in Exons (56,39 %) am niedrigsten ist. Bei der Analyse der Nukleotidzusammensetzung in der ersten 5'UTR fällt auf, dass der AT-Gehalt mit 61,89 % in etwa dem der gesamten 5'UTR entspricht. Bei den fünf ersten Introns der Gene befindet sich der AT-Gehalt in derselben Größenordnung, auch wenn dieser mit zunehmender Anzahl an Introns leicht abnimmt (von 68,13 % im ersten Intron auf 67,18 % im fünften Intron). Bei den vier ersten Exons der Gene ist der AT-Gehalt sogar noch konstanter als dies bei den Introns der Fall ist. Lediglich bei den Nachkommastellen tritt eine Änderung auf (von 56,1 % bis 56,47 %). Der Vergleich zwischen

der gesamten 3'UTR (66,47 %) und der ersten 3'UTR (66,71 %), tritt ebenfalls nur in den Nachkommastellen ein veränderter AT-Gehalt auf.

Des Weiteren wurde die Anzahl der Nukleotide bestimmt, die im intergenischen Bereich und innerhalb von Genen liegen. Die Gesamtlänge der DNA auf den Chromosomen beträgt 119186497 Nukleotide. 43,7 % der Nukleotide befinden sich im intergenischen Bereich, während 56,3 % innerhalb von Genen lokalisiert sind. Die Anzahl der Nukleotide in der 5'UTR beträgt 1,78 %, in Introns 15,86 %, in Exons 35,18 % und in der 3'UTR 3,50 %. Die Bestimmung der Anzahl der Nukleotide innerhalb der ersten 5'UTR, der ersten fünf Introns, der ersten fünf Exons und der ersten 3'UTR wurde nicht durchgeführt.

Die Verteilungsanalysen von Transkriptionsfaktor-Bindungsstellen wurden im intergenischen Bereich sowie innerhalb der Gene durchgeführt. In einigen Fällen wurde hierbei nicht z.B. das gesamte erste Intron berücksichtigt, sondern nur die durchschnittliche Nukleotidlänge. Daher wurden bei der Analyse der Nukleotidkomposition nun die durchschnittlichen Längen der ersten 5'UTR, der ersten fünf Introns, der ersten vier Exons sowie der ersten 3'UTR untersucht. Weiterhin wurde der durchschnittliche AT-Gehalt jeder einzelnen Position in den ausgewählten Bereichen dargestellt. Zusätzlich wurde die Analyse auch auf den jeweils am höchsten nummerierten Bereich der Gene ausgeweitet. Die Positionsangaben der höchsten Bereiche wurden nicht wie bei den anderen durch SQL-Abfragen, sondern mit den TFBSitesTools (Kapitel 2.5.3) ermittelt. Insgesamt wurde die Analyse jeweils vom Beginn eines Bereichs sowie vom Ende eines Bereichs aus durchgeführt.

3.1.1 5'UTR

Nachdem der AT-Gehalt für die einzelnen Bereiche bestimmt wurde, erfolgte die detaillierte Betrachtung des AT-Gehalts innerhalb der durchschnittlichen Länge der ersten 5'UTR, vom Beginn der 5'UTR aus betrachtet. Das heißt, mit zunehmender Distanz wird der Abstand zum Translationsstart immer geringer. Wie in Abb. 12 dargestellt variiert der AT-Gehalt in den ersten Nukleotiden am Anfang der 5'UTR stärker als dies im weiteren Verlauf der Fall ist. Abgesehen von den ersten Nukleotiden beträgt der AT-Gehalt durchschnittlich knapp über 60 %, was eine gute Korrelation zu dem durchschnittlichen AT-Gehalt der ersten 5'UTR darstellt.

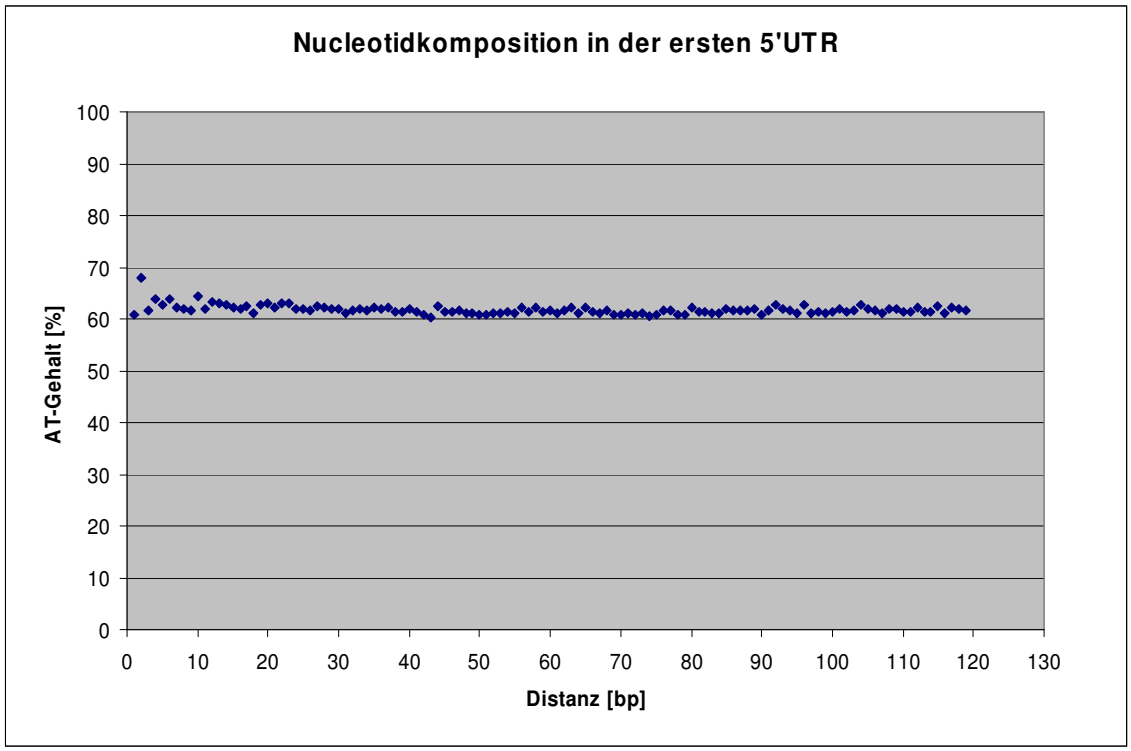


Abb. 12: AT-Gehalt der ersten 5'UTR vom Beginn des Bereichs aus bis zur durchschnittlichen Länge von 119 bp betrachtet. Bis auf die Anfangsnukleotide befindet sich der AT-Gehalt in derselben Größenordnung.

Da zu Beginn der ersten 5'UTR der AT-Gehalt stärker variiert als dies im weiteren Verlauf der Fall ist, wurden zur detaillierten Ansicht die ersten 40 Positionen gesondert grafisch aufgetragen (Abb. 13). Es wurde jeweils die durchschnittliche Anzahl der verschiedenen Nukleotide an jeder Position grafisch dargestellt.

Zu Beginn der ersten 5'UTR ist ein vermehrtes Auftreten des Nukleotids A zu verzeichnen (50,62 %). Innerhalb der gesamten betrachteten Nukleotidlänge ist die Sequenz der ersten 5'UTR AT-reich. Bis auf die ersten Nukleotide treten diese für sich betrachtet während der gesamten betrachteten Länge gleich häufig auf (C: 22 % bis 23 %). Das Nukleotid C kommt bis auf den Beginn der ersten 5'UTR häufiger vor als G.

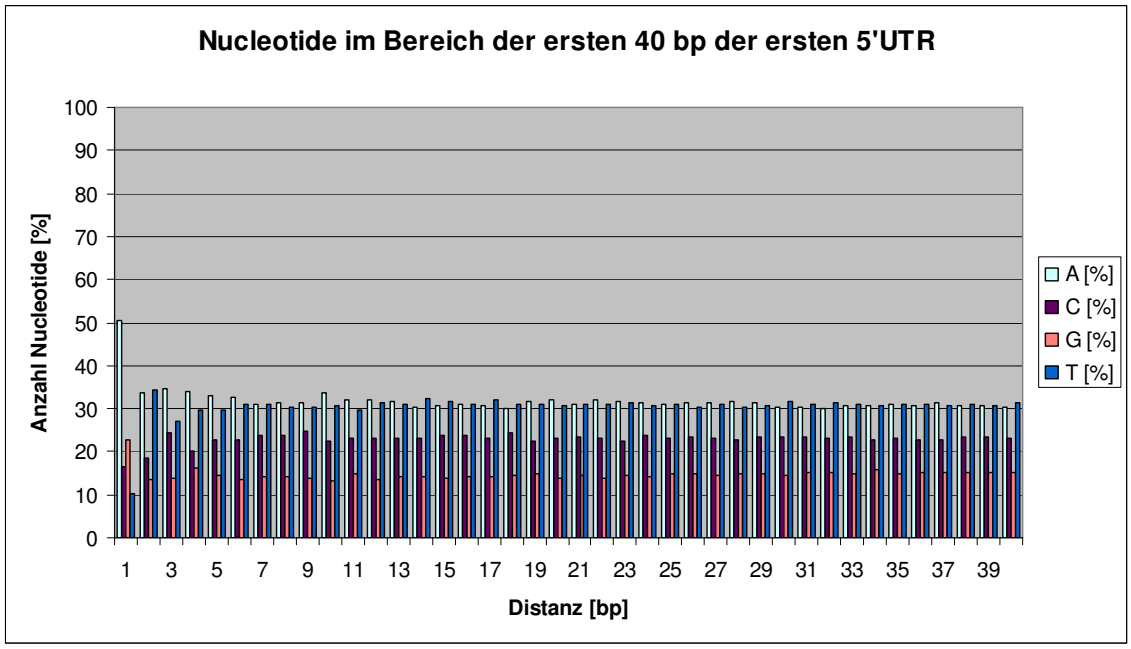


Abb. 13: Prozentuale Nukleotidanzahl in den ersten 40 Positionen der ersten 5'UTR betrachtet vom Beginn der 5'UTR aus. Am Anfang der 5'UTR unterscheidet sich die Nukleotidverteilung vom Rest des betrachteten Bereichs.

Um den Informationsgehalt der betrachteten Sequenzen einschätzen zu können, wurde ein Sequenzlogo (Kapitel 2.8) mit den ersten 40 Nukleotiden erstellt (Abb. 14). Die erste Position der ersten 5'UTR ist konservierter als die restlichen analysierten Positionen. Bevorzugt wird an der ersten Position das Nukleotid A. Im Vergleich zu anderen betrachteten Sequenzlogos (z.B. Abb. 20 und Abb. 23) sind die bevorzugten Nukleotide an einigen weiteren Positionen weniger zufällig, jedoch ist der Informationsgehalt nicht hoch genug, um eine Aussage über eine reale Bevorzugung von Nukleotiden zu treffen. Nach der 27. Position ist keinerlei Bevorzugung von Nukleotiden detektierbar.

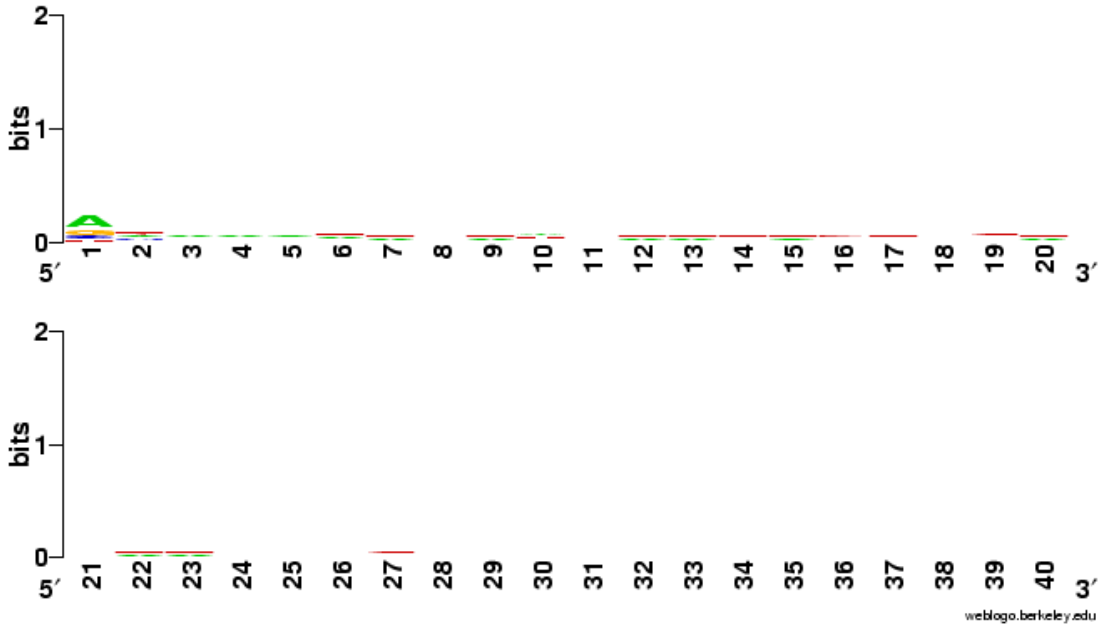


Abb. 14: Sequenzlogo der ersten 40 Positionen der ersten 5'UTR. Nur an der ersten Position ist eine Bevorzugung für ein Nukleotid (A) erkennbar. Die restlichen Positionen zeigen keine Bevorzugungen.

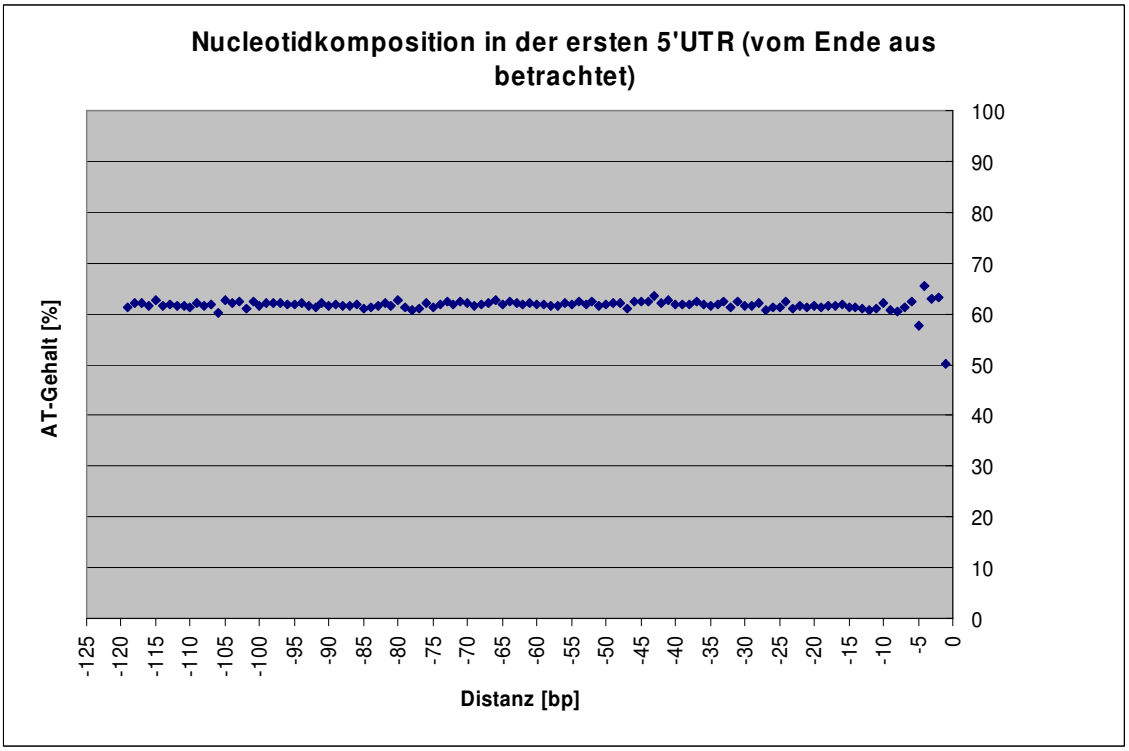


Abb. 15: AT-Gehalt der ersten 5'UTR von der durchschnittlichen Länge bis zum Ende des Bereichs aus betrachtet. Abgesehen von den Nukleotiden am Ende der 5'UTR befindet sich der AT-Gehalt in derselben Größenordnung.

Des Weiteren wurde eine Analyse der Nukleotidkomposition vom Ende der 5'UTR aus durchgeführt. Das Ende der 5'UTR liegt dem Translationsstartpunkt am nächsten. Abb. 15 zeigt den AT-Gehalt der, vom Ende der ersten 5'UTR aus gesehen, ausgerichteten Sequenzen. Generell ist der AT-Gehalt vergleichbar mit dem der ersten 5'UTR vom Anfang aus betrachtet, was zu erwarten ist. Die letzten Positionen, welche die sind, die dem Translationsstartpunkt am nächsten sind, weichen jedoch im AT-Gehalt von den übrigen ab. Daher werden diese 40 Positionen gesondert betrachtet (Abb. 16).

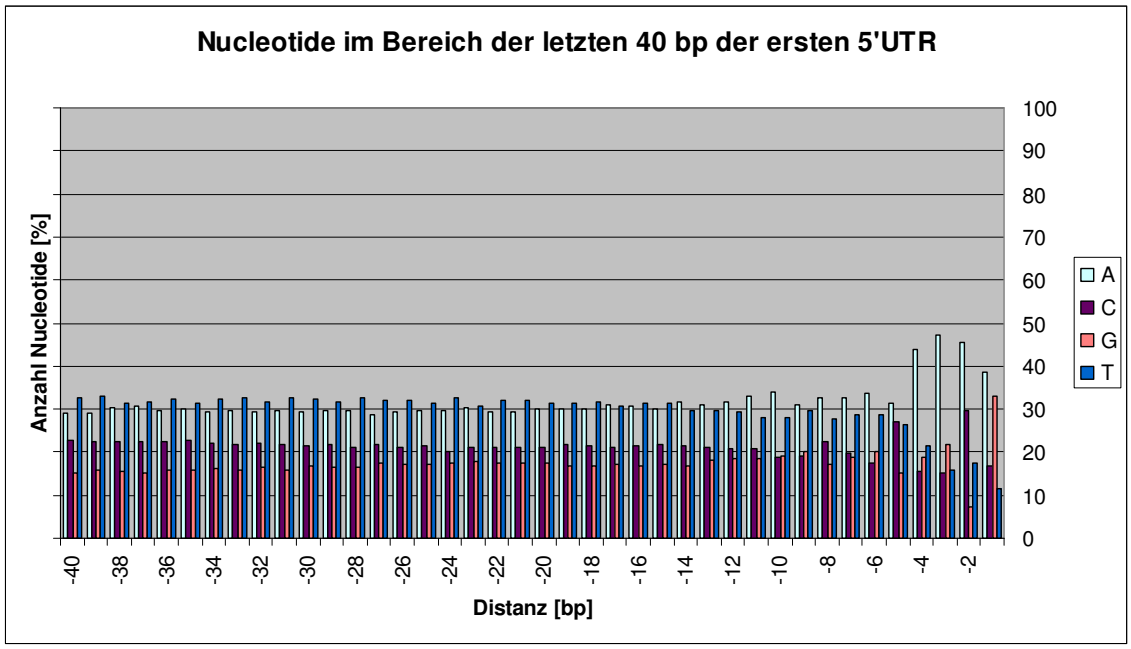


Abb. 16: AT-Gehalt der letzten 40 Positionen der ersten 5'UTR. Abgesehen von den Nukleotiden am Ende der 5'UTR, deren letzten 4 Positionen sich in der Nukleotidkomposition von den restlichen deutlich unterscheiden, befindet sich der AT-Gehalt in derselben Größenordnung, Erläuterungen siehe Text.

Die letzten 4 Positionen der ersten 5'UTR unterscheiden sich von den vom Translationsstartpunkt weiter entfernt gelegenen. Zunächst wird das Nukleotid A bevorzugt, wobei dieses an den Positionen -2 und -3 zunimmt und an der Position -4 leicht wieder abnimmt. An der Position -1 tritt alternativ das Nukleotid G auf. Auch bei der Betrachtung vom Ende der 5'UTR aus gesehen treten die vier Nukleotide jeweils etwa gleich häufig innerhalb der betrachteten Sequenzlänge auf.

In dem entsprechenden Sequenzlogo dieser Region, dargestellt in Abb. 17, kann die Bevorzugung für das Nukleotid A am Ende des Bereichs bestätigt werden, wobei der Informationsgehalt an der Position -2 seinen Maximalwert erreicht. Lediglich die letzten vier

Positionen weisen einen darstellbaren Informationsgehalt auf, das deutet darauf hin, dass die vorangegangenen Positionen nicht konserviert sind.

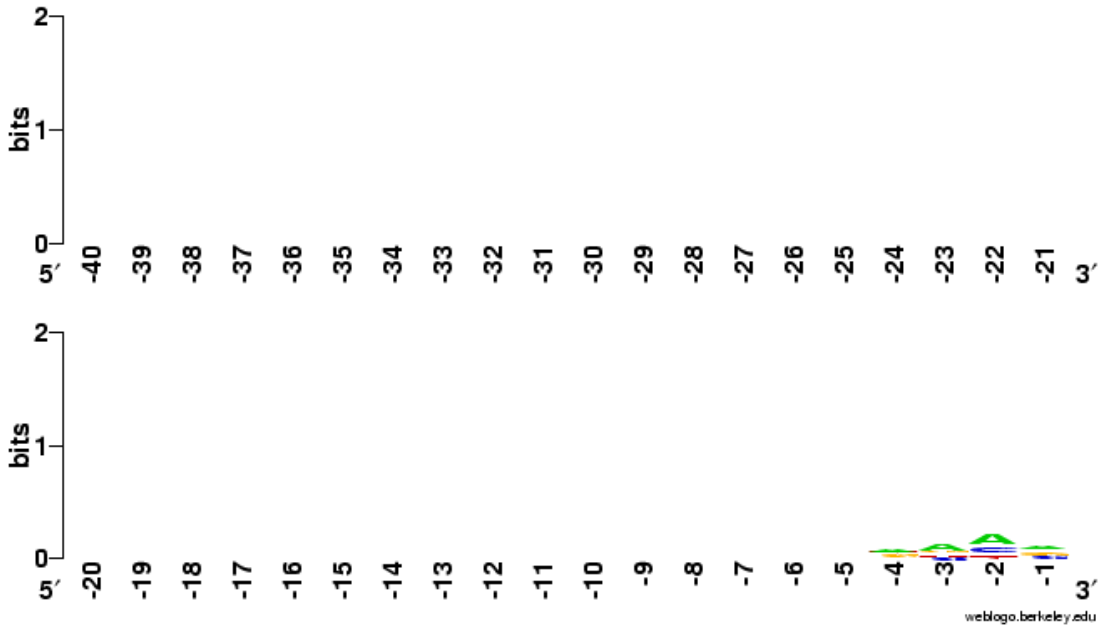


Abb. 17: Sequenzlogo der letzten 40 Positionen der ersten 5'UTR. Der Informationsgehalt der Sequenzen von Position -5 bis -40 ist zu gering, um dargestellt werden. Die letzten 4 Positionen, das Ende der ersten 5'UTR, weisen eine Bevorzugung für das Nukleotid A auf.

Prinzipiell können bei der Untersuchung der ersten 5'UTR auch noch ein oder mehrere Introns in der 5'UTR liegen. Daher wurde die dem Translationsstart am nächsten liegende 5'UTR, welche die am höchsten nummerierte ist, zur Kontrolle ebenfalls analysiert, wobei diese ebenfalls vom Ende aus betrachtet wurde. Tendenziell unterscheidet sich die Nukleotidverteilung nicht von der der ersten 5'UTR, auch das entsprechende Sequenzlogo weist keine wesentlichen Unterschiede auf (Daten nicht gezeigt). Dies deckt sich mit der Erkenntnis, dass bei den meisten Genen die 5'UTR nicht durch ein Intron unterbrochen wird.

3.1.2 Introns

Analog zu den Analysen betreffend der 5'UTR erfolgt nun die Betrachtung der Nukleotidkomposition der ersten fünf Introns sowie der am höchsten nummerierten Introns der jeweiligen Gene. Da die Ergebnisse für die einzelnen Introns weitestgehend übereinstimmen, werden im Folgenden nur die Ergebnisse für das erste Intron grafisch dargestellt.

Werden die Nukleotide innerhalb der durchschnittlichen Länge der Introns vom Beginn des Introns aus betrachtet, ist bei den ersten etwa 20 Positionen der Introns im Vergleich zu den

restlichen Positionen ein variierender AT-Gehalt zu verzeichnen (Abb. 18). Im weiteren Verlauf liegt der AT-Gehalt bei knapp unter 70 %. Introns beginnen in der Regel mit GT und enden mit AG (alternative Introns: AT-AC-Introns, Meli *et al.*, 2001)), daher überrascht es nicht, dass die erste Position nur zu 0,03 % ein A oder T beinhaltet, während an der zweiten Position zu 99,46 % ein A oder T auftritt.

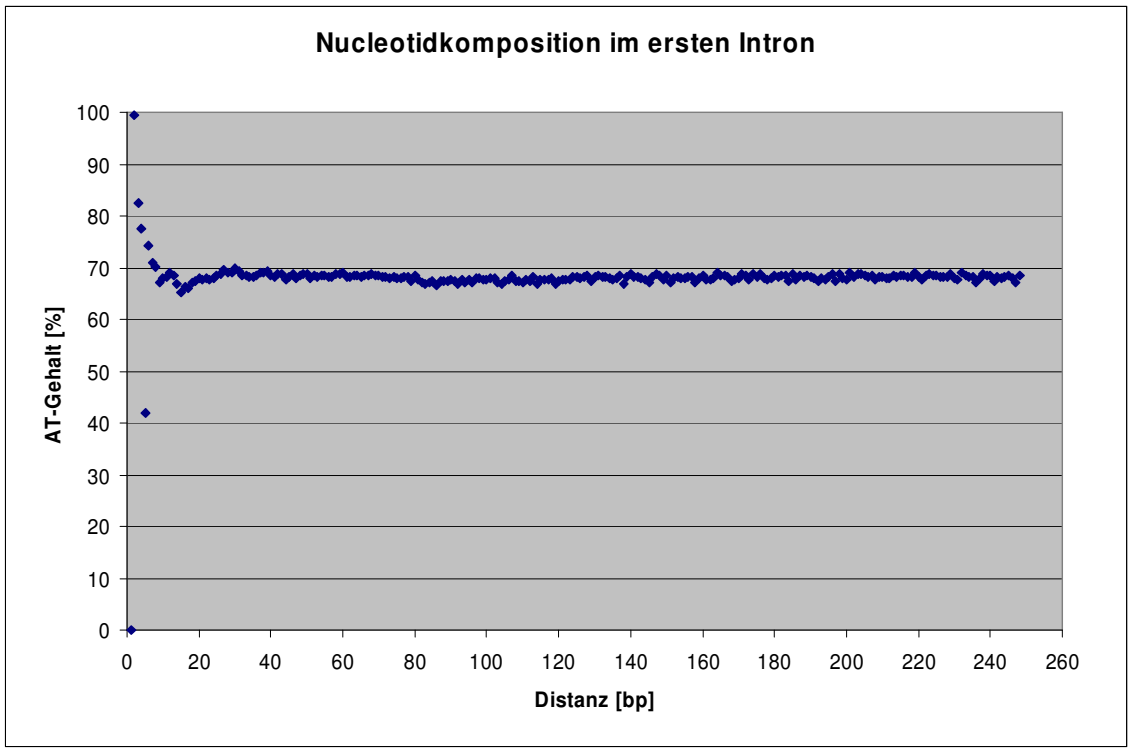


Abb. 18: AT-Gehalt innerhalb der durchschnittlichen Nukleotidlänge (248 bp) des ersten Introns vom Beginn aus betrachtet. Die anfänglichen Positionen weisen einen unterschiedlichen AT-Gehalt im Vergleich zu den übrigen Positionen auf.

Um den Beginn der Introns näher zu betrachten, wurden die ersten 40 Positionen grafisch aufgetragen (Abb. 19). Fast 100 % der Introns beginnen mit GT, danach schließt sich eine Bevorzugung von AA an den Positionen 3 und 4 an. Im weiteren Verlauf tritt das Nukleotid T häufiger als die anderen Nukleotide auf. Die Nukleotidverteilung ist leicht wellenförmig. Wird diese ab der Position 5 genauer betrachtet, weist T zwei Minima bei den Positionen 15 und 24 und A bei 5 und 15 auf. Das Nukleotid C zeigt diese Minima bei 5 und 12 und ist danach abflachend verteilt, während G einen Minimalwert bei 8 besitzt (Abb. 19). Der wellenförmige Verlauf der Nukleotide ist bei allen betrachteten Introns zu verzeichnen, teilweise sind die Minimalwerte bei den Nukleotiden A (Positonen 5 und 10, wobei im ersten Intron an der Position 10 kein Minimum aufweisbar ist) und C (Positionen 5 und 12) in den einzelnen

Introns identisch (Daten nicht gezeigt). Das Nukleotid G weist bei allen Introns ein Minimum an der Position 6 oder 8 auf.

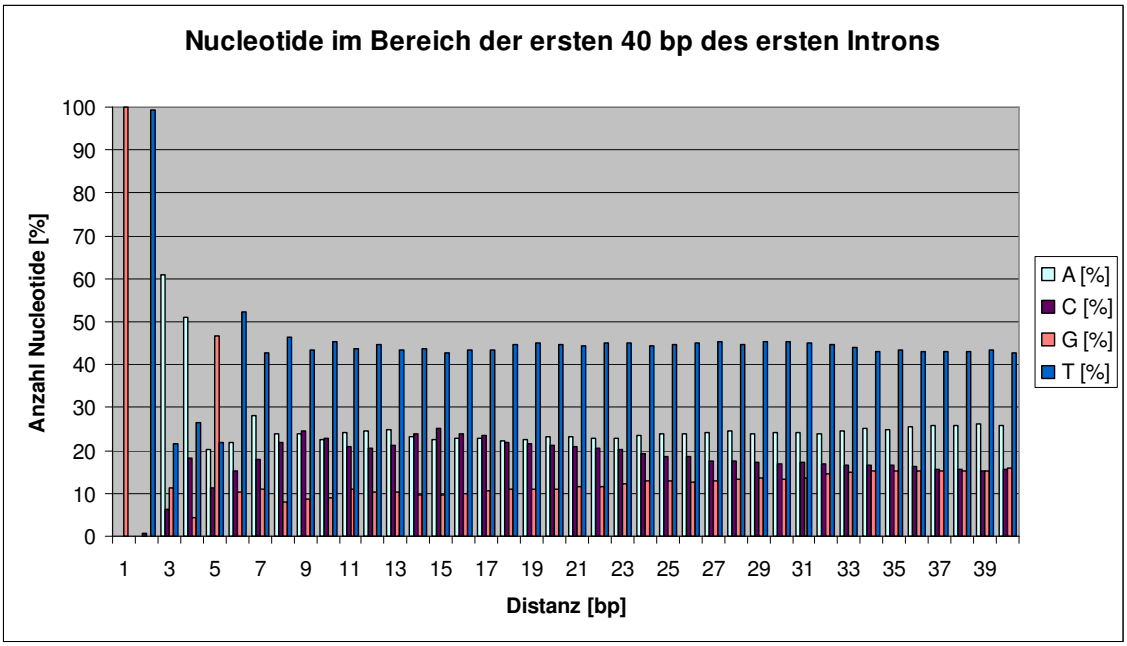


Abb. 19: Prozentuale Anzahl der ersten 40 Nukleotide vom Beginn des ersten Introns aus betrachtet. Introns beginnen fast ausschließlich mit GT, anschließend ist eine Präferenz für AAGT zu verzeichnen. Das Nukleotid T ist an den weiteren Positionen vorwiegend vertreten.

Die Bevorzugung der Anfangsnukleotide GT zu Beginn des ersten Introns lässt sich auch im entsprechenden Sequenzlogo (Abb. 20) bestätigen. Dies deutet auf eine fast identische Nukleotidkomposition zu Beginn von Introns hin. Diese tritt sowohl bei allen betrachteten ersten fünf Introns der Gene sowie bei den höchsten Introns der Gene auf. Im Sequenzlogo wird ebenfalls das nachfolgende bevorzugte Auftreten von AA widergespiegelt, wenn auch diese beiden Nukleotide einen geringeren Informationsgehalt aufweisen. Im weiteren Verlauf wird das präferenzielle Vorkommen des Nukleotids T bestätigt.

In der Folge wurden die fünf ersten Introns sowie das am höchsten nummerierte Intron der Gene vom Ende aus gesehen analysiert. Da auch hier, wie bei der Betrachtung vom Beginn der Introns aus, bei den ersten fünf untersuchten Introns der Gene keine wesentlichen Unterschiede aufzufinden waren, werden die Grafiken exemplarisch für das erste Intron angeführt.

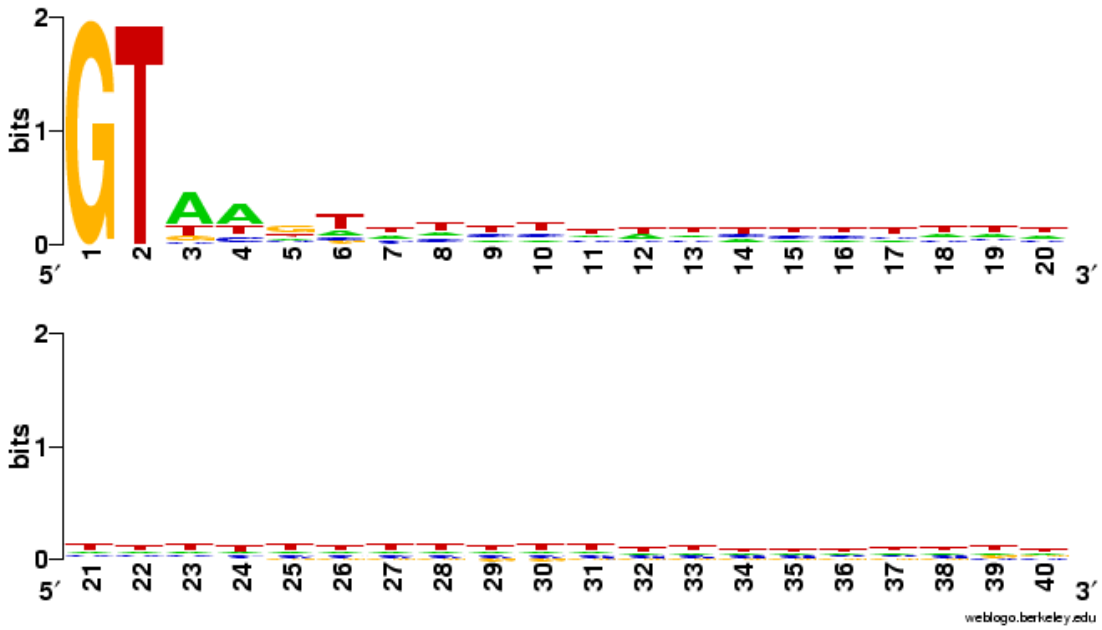


Abb. 20: Sequenzlogo der ersten 40 Positionen des ersten Introns vom 5' Start aus betrachtet. Introns beginnen bevorzugt mit GT, wonach sich eine präferenzielle Anwesenheit von AA(G)T anschließt.

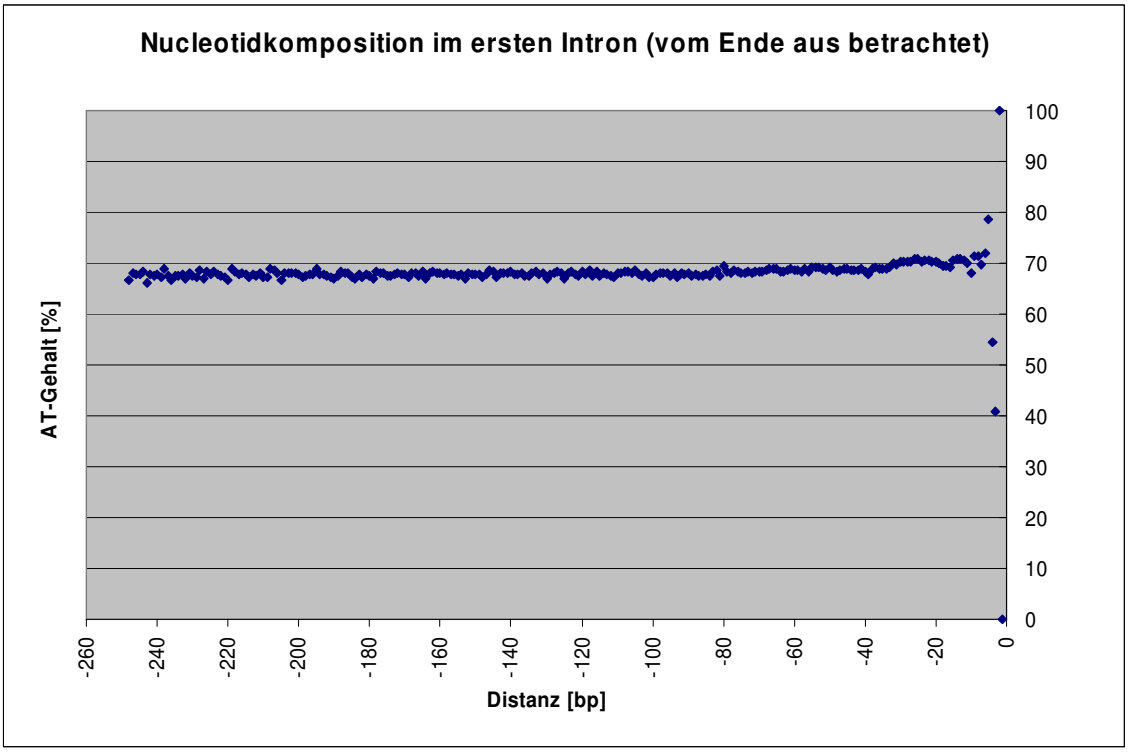


Abb. 21: AT-Gehalt des ersten Introns der durchschnittlichen Länge (248 bp) vom Ende aus betrachtet, das heißt, die Sequenzen wurden am Endpunkt aliniert. Der AT-Gehalt liegt bei knapp unter 70 % und weist eine starke Variation am Ende des Introns auf.

Bei der Betrachtung der Introns vom Ende aus gesehen lässt sich bei der Darstellung des AT-Gehaltes zunächst am Ende eine Streuung feststellen, während mit zunehmender Distanz vom Ende aus gesehen der AT-Gehalt konstant wirkt und sich in einer Größenordnung von knapp unter 70 % bewegt (Abb. 21). Dies stimmt mit der Beobachtung vom Beginn des Introns aus gesehen überein, was darauf hinweist, dass der AT-Gehalt zwischen dem Anfang und dem Ende des Introns bei einem etwa konstanten Wert liegt und sich möglicherweise unter den Analysebedingungen nicht wesentlich ändert.

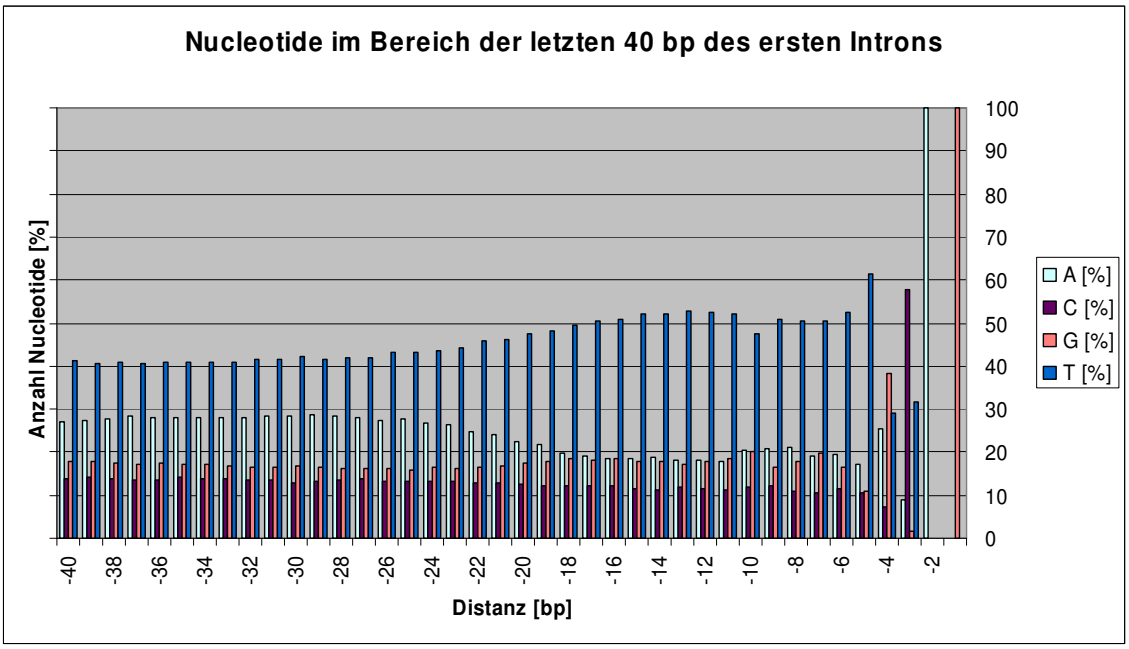


Abb. 22: Nukleotiddarstellung der letzten 40 Positionen des ersten Introns. Zum Ende hin nimmt der Gehalt an T wellenförmig zu, während das Auftreten von A abnimmt. Introns enden bevorzugt mit T(G)CAG.

Zur besseren Darstellung des direkten Endes der Introns wurde die prozentuale Anzahl der jeweiligen Nukleotide in den letzten 40 bp dargestellt (Abb. 22). Der wellenförmige Charakter der Nukleotidverteilung tritt bei allen betrachteten Introns auf. Auffällig ist, dass sich A und T komplementär zueinander verhalten, nimmt T zu, nimmt A zugleich ab und umgekehrt. Dies deutet auf eine gemeinsame Lokalisierung von bestimmten Nukleotiden an Positionen am Ende von Introns hin. Introns werden in der Regel mit AG beendet, was durch die Grafik bestätigt wird. An der Position -3 tritt vorzugsweise ein C auf. An der Position -5 liegt eine starke Präferenz von mehr als 60 % für das Nukleotid T vor, wobei an den nachfolgenden Positionen ebenfalls eine Bevorzugung für das Nukleotid T auftritt, welche wellenförmig ab- und zu- und abnimmt, bis der Wert ab etwa -27 bp einen konstanten Wert anzunehmen scheint. Diese Beobachtungen lassen sich auch im entsprechenden Sequenzlogo bestätigen

(Abb. 23). Im Bereich von -40 bis -5 Positionen bis zum Ende des Introns tritt das Nukleotid T häufiger als andere Nukleotide auf und nimmt mit abnehmender Distanz zu. Am Ende des Introns kommt es zu einer Bevorzugung von (T)N(C/T)AG, während fast alle Introns mit AG aufhören.

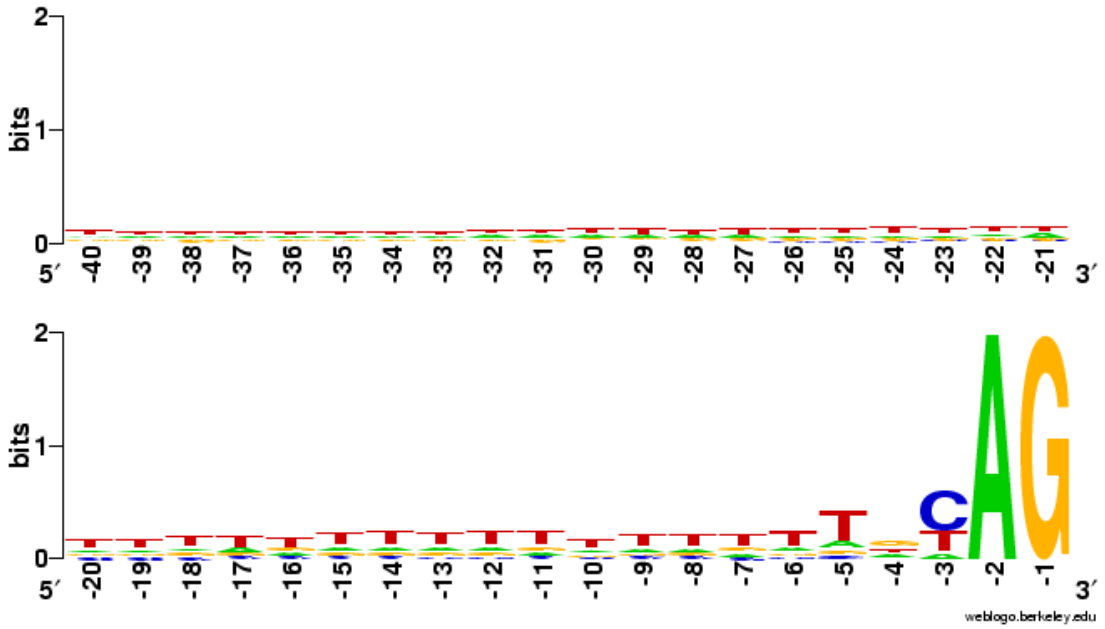


Abb. 23: Sequenzlogo der letzten 40 Positionen des ersten Introns von Genen. Das Nukleotid T tritt bevorzugt am Ende der Introns auf, wobei die Präferenz mit abnehmender Distanz zum Ende des Introns zunimmt. Das erste Intron von Genen endet bevorzugt mit (T)N(C/T)AG.

Auch dieses Sequenzlogo stimmt mit denen für die anderen betrachteten Introns überein. Das legt den Schluss nah, dass es sich um eine bevorzugte Nukleotidkomposition am Ende von Introns handeln könnte. Insgesamt beginnen 121 erste Introns nicht mit einem GT und 12 Introns enden nicht mit AG.

3.1.3 Exons

Nach der Analyse der 5'UTR und der Introns wurden die vier ersten Exons sowie das letzte Exon jedes Gens untersucht. Wie bei den vorangegangenen Analysen wurden sowohl der Beginn als auch auch das Ende der Exons betrachtet. Als Beginn des ersten Exons wird der Translationsstartpunkt angesehen, im Gegensatz zur TIGR Annotation (Kapitel 2.1), bei der es sich bei einem Exon um einen transkribierten Bereich handelt.

Der durchschnittliche Bereich des ersten Exons ist sehr viel größer als dies bei den darauffolgenden Exons der Fall ist (Kapitel 2.1.1). Das hängt vermutlich damit zusammen, dass einige Gene nur ein Exon aufweisen, welches länger sein könnte als Exons bei Genen,

die Introns besitzen. Abb. 24 stellt den AT-Gehalt im ersten Exon auf der durchschnittlichen Länge von 692 Nukleotiden dar. Zu Beginn des Exons variiert der AT-Gehalt stark, was vor allem dadurch begründet ist, dass das erste Exon in der Regel mit einem ATG beginnt, welches den Translationsstartpunkt beschreibt. Nachfolgend kommt es zu einer Streuung des AT-Gehalts im gesamten betrachteten Bereich. Diese Streuung ist in einem Bereich von etwa 50 % bis 60 % lokalisiert und erscheint in 3 Basen Intervallen. Wahrscheinlich ist diese durch die Codon Usage begründet (http://www.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=3702). Laut dieser Quelle beträgt der Durchschnitt des AT-Gehalt bei *Arabidopsis thaliana* Exons der jeweiligen Codons 55,41 % an der Position 1, 49,16 % für die Position 2 sowie an dritter Position 57.62 %.

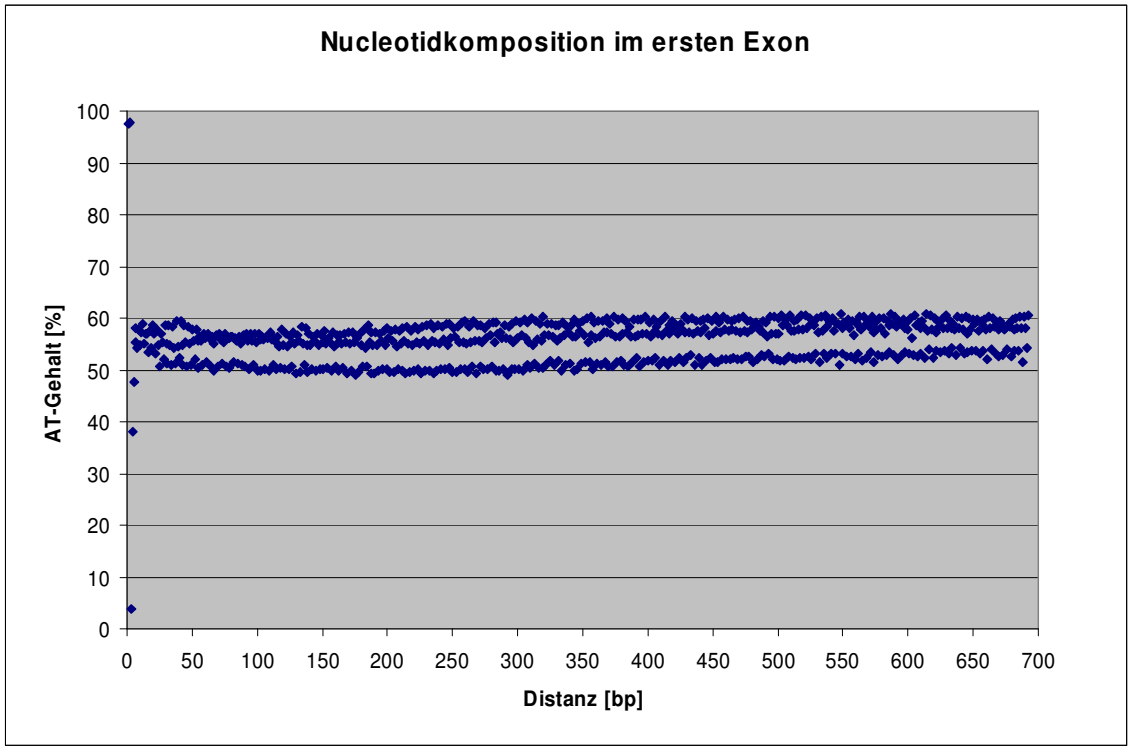


Abb. 24: AT-Gehalt der durchschnittlichen Länge von 692 bp des ersten Exons der Gene. Zu Beginn kommt es zu einer starken Abweichung des AT-Gehalts im Vergleich zu weiter entfernten Distanzen. Generell ist eine Streuung des AT-Gehaltes zu verzeichnen, die sich über den gesamten Bereich erstreckt und zum Ende hin weniger deutlich zum Vorschein tritt.

Um einen detaillierteren Eindruck über die Nukleotidkomposition am Anfang des Exons zu erhalten, wurden die ersten 40 Positionen gesondert aufgetragen (Abb. 25). Deutlich wird, dass nicht alle Translationsstartpunkte durch ein ATG determiniert werden, da die Häufigkeit von A an der ersten Position nicht 100 % beträgt. Des Weiteren zeigt die Abbildung, dass die vierte und fünfte Position deutlich geringere AT-Gehalte aufweisen als dies bei weiter

entfernt liegenden Positionen der Fall ist. Die vierte Position zeigt eine Bevorzugung für das Nukleotid G. Im weiteren Verlauf ist keine eindeutige Präferenz für ein Nukleotid zu beobachten.

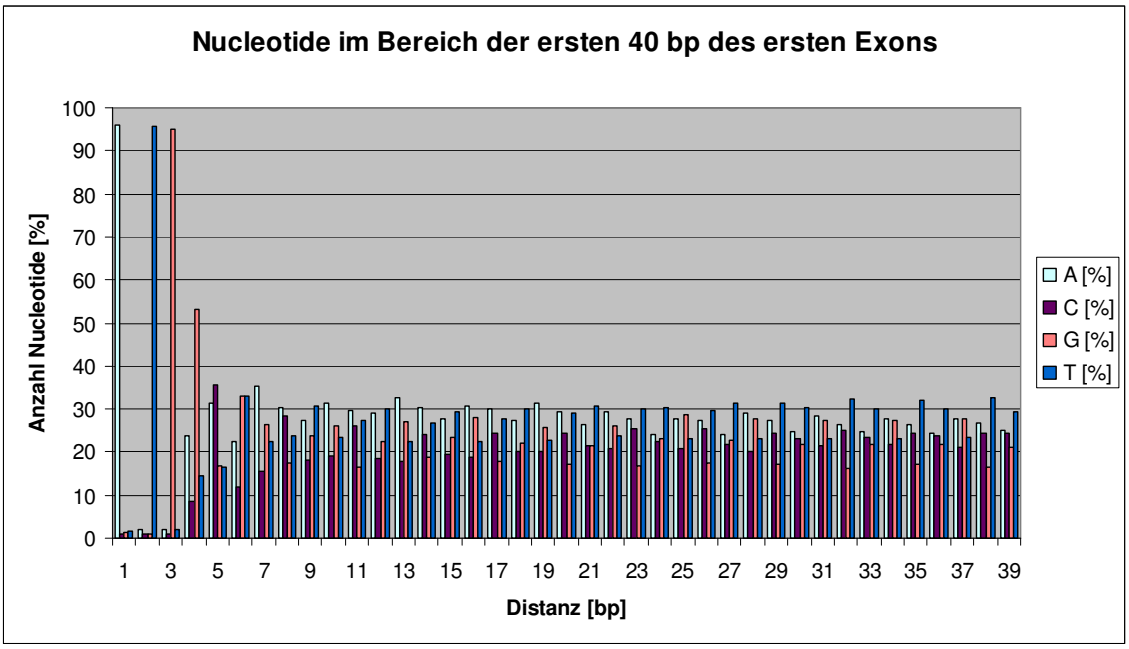


Abb. 25: Darstellung der ersten 40 Positionen des ersten Exons der jeweiligen Gene. Zu Beginn des Exons tritt eine Bevorzugung von ATG auf, danach ist eine Präferenz für das Nukleotid G zu verzeichnen. Im weiteren Verlauf tritt keine eindeutige Bevorzugung für Nukleotide an bestimmten Positionen auf.

Um den Informationsgehalt und damit den Einfluss des Zufalls auf die erhaltenen Ergebnisse zu analysieren, wurde ein Sequenzlogo mit den ersten 40 Positionen durchgeführt. Da das Sequenzlogo des Beginns des ersten Exon mit allen vorhandenen Sequenzen aufgrund einer Limitierung der Anzahl der bei WebLogo eingebbaren Sequenzen fehlschlug, wurden für die Erzeugung nur die ersten 24000 (von 29993) Sequenzen verwendet (Abb. 26). Das Sequenzlogo bestätigt die starke Präsenz des ATGs zu Beginn des ersten Exons, und auch die Präferenz für das G an der vierten Position lässt sich bestätigen. Die weiteren Positionen zeigen aufgrund des Informationsgehaltes keine darstellbare Präferenz.

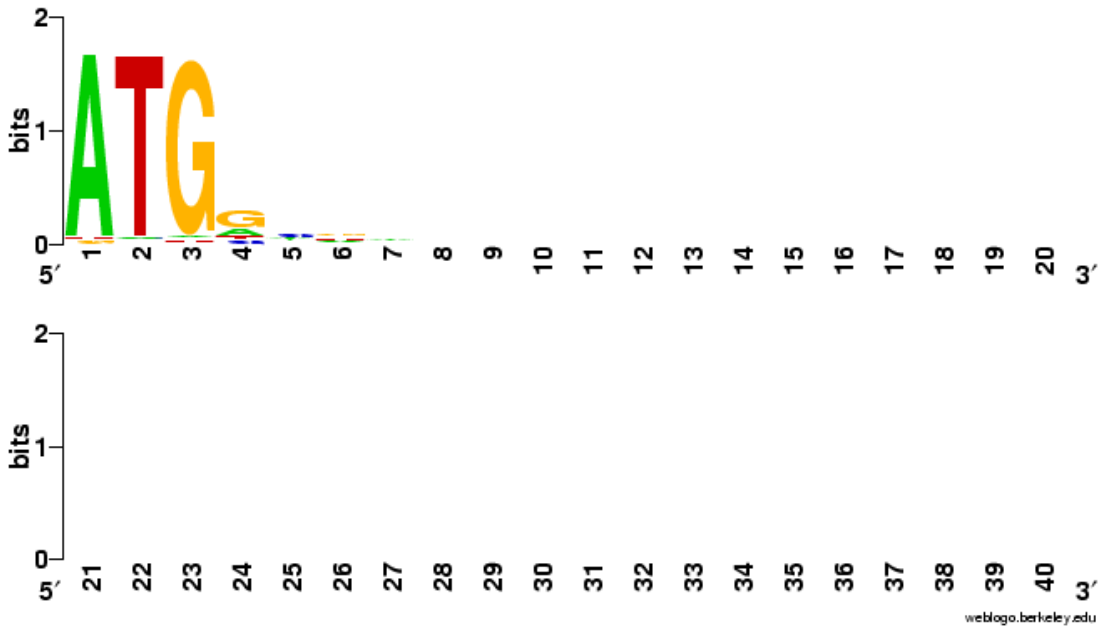


Abb. 26: Sequenzlogo mit den ersten 40 Positionen von 24000 Sequenzen des erstes Exons der jeweiligen Gene. Das ATG als bevorzugter Translationsstart tritt bevorzugt zum Beginn des ersten Exons auf. Danach schließt sich eine leichte Bevorzugung des Nukleotids G an. Ab der Position 8 ist keinerlei Informationsgehalt darstellbar, daher kann davon ausgegangen werden, dass keine weitere systematische Bevorzugung von bestimmten Nukleotiden unter den Analysebedingungen auftritt.

Im Vergleich zum ersten Exon unterscheidet sich die Nukleotidverteilung in den folgenden drei Exons. Die Exons 2 bis 4 unterscheiden sich in der Nukleotidverteilung nicht voneinander. Beispielhaft sind die Ergebnisse für das zweite Exon dargestellt.

Nachdem der AT-Gehalt zu Beginn stärker streut, pendelt er sich dann in einem Bereich zwischen 50 % und 60 % ein (Abb. 27). Werden die ersten 40 Positionen gesondert betrachtet, ist zu erkennen, dass nur an den ersten beiden Positionen eine deutliche Bevorzugung für ein bestimmtes Nukleotid auftritt, C und T (Daten nicht gezeigt). Bei den restlichen Positionen kommen die vier Nukleotide für sich betrachtet etwa ähnlich häufig vor, jedoch tritt das Nukleotid C weniger häufig auf als die anderen drei.

Mit den ersten 40 Positionen der zweiten Exons von Genen wurde ein Sequenzlogo erstellt (Abb. 28). Das Sequenzlogo spiegelt die leichte Bevorzugung von G und T an den ersten beiden Positionen wieder, ab der Position 4 aufgrund des Informationsgehalts keine weitere Präferenz darstellbar. Gleiche Ergebnisse konnten für die Exons 3 und 4 ermittelt werden (Daten nicht gezeigt). Bei der Untersuchung des höchsten Exons kam es bei der Analyse des Anfangsbereichs zu einer Überlagerung des ATGs (Abb. 26) und des GTs (Abb. 28), was darauf zurückzuführen ist, dass einige Gene nur ein Exon aufweisen. Auf die Darstellung des höchsten Exons wird deshalb verzichtet.

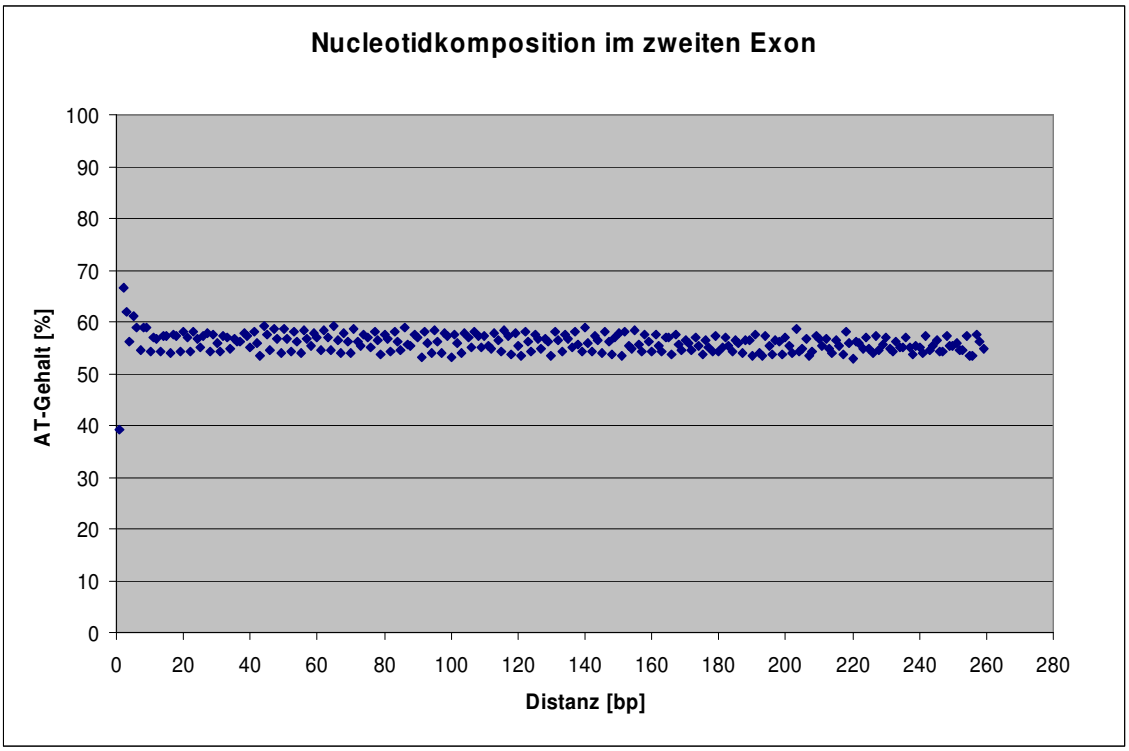


Abb. 27: AT-Gehalt innerhalb der durchschnittlichen Länge von 259 bp des zweiten Exons der jeweiligen Gene. Zu Beginn tritt eine Variation des AT-Gehalts im Vergleich zu den übrigen Positionen auf, der restliche Verlauf zeigt eine Streuung des AT-Gehalts.

Abb. 28: Sequenzlogo der ersten 40 Positionen der zweiten Exons der jeweiligen Gene. Zu Beginn treten präferenziell die Nukleotide GT auf, ab der Position 4 ist aufgrund des Informationsgehalts keine Präferenz mehr darstellbar. Daher kann davon ausgegangen werden, dass es in dem betrachteten

Bereich zu keiner weiteren biologisch relevanten Bevorzugung von Nukleotiden an bestimmten Positionen kommt.

Bei der Betrachtung der Nukleotidkomposition vom Ende der Exons aus wurden beim 2. bis 4. Exon vergleichbare Ergebnisse erzielt. Beispielhaft sind diese für das Exon 2 dargestellt (Abb. 29). Nachdem mit zunehmender Distanz vom Ende aus gesehen der AT-Gehalt zwischen 50 % und 60 % liegt, verändert er sich zum Ende hin, wobei er zunächst in den letzten 20 Nukleotiden ansteigt, danach seine Extremwerte erreicht und an der letzten Position rapide abfällt. Die letzten 40 Positionen wurden gesondert betrachtet und weisen bis auf die letzten Positionen keine Bevorzugung für bestimmte Nukleotide auf, diese treten ähnlich häufig auf (Daten nicht gezeigt).

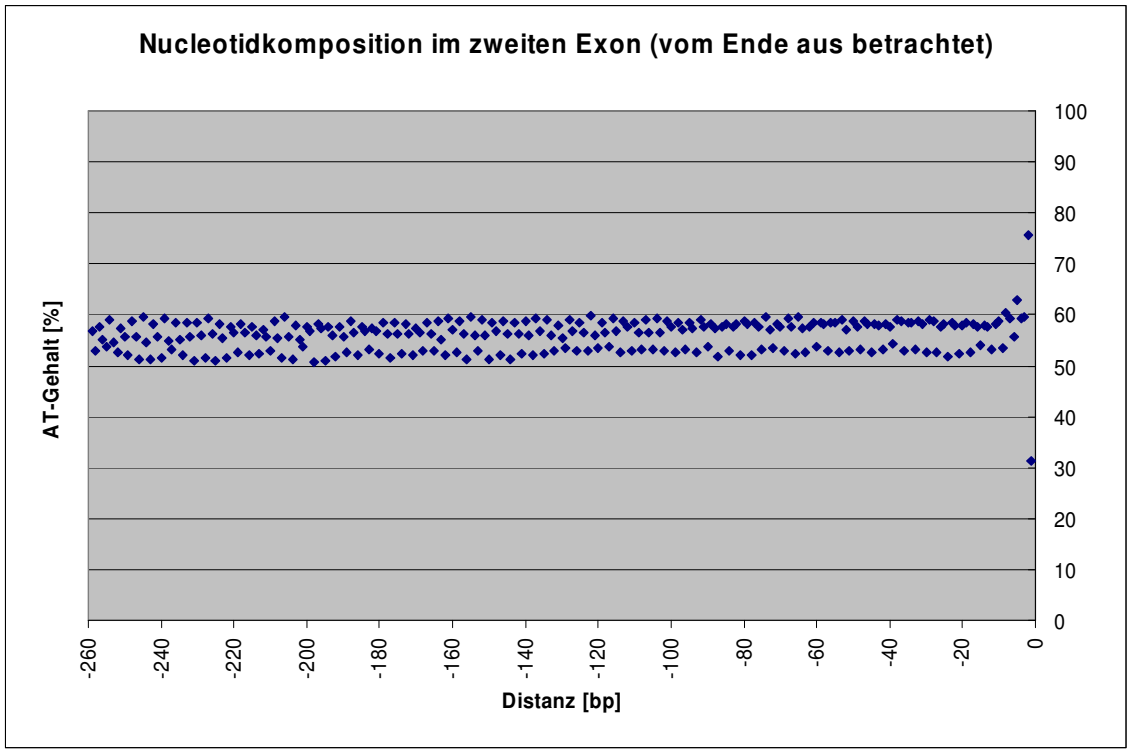


Abb. 29: Darstellung des AT-Gehalts des zweiten Exons auf einer durchschnittlichen Länge von 259 Nukleotiden, wobei die Sequenzen am Ende des Exons ausgerichtet wurden. Es kommt zu einer Streuung des AT-Gehalts innerhalb von 50 % und 60 % bis kurz vor dem Ende des Exons. Hier steigt der AT-Gehalt tendenziell an und erreicht seine größte Abweichung an den letzten beiden Positionen.

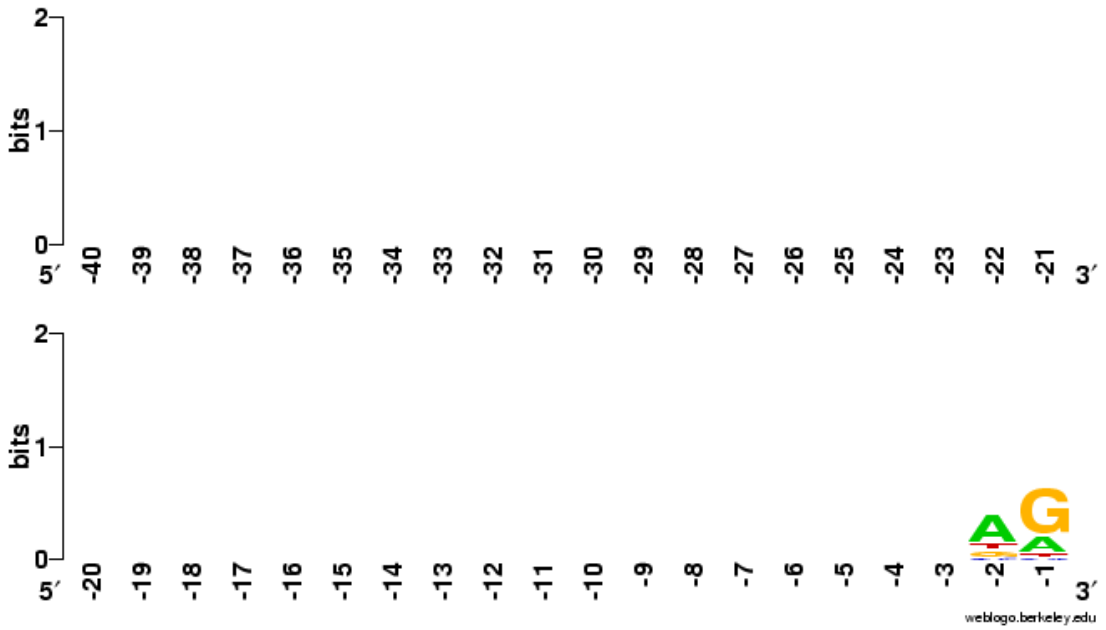


Abb. 30: Sequenzlogo der letzten 40 Positionen des zweiten Exons vom Ende der Exons aus betrachtet. Bis auf die letzten beiden Positionen ist aufgrund des Informationsgehalts keine weitere Präferenz darstellbar, was darauf hindeutet, dass es generell zu keiner weiteren Bevorzugung von bestimmten Nukleotiden an bestimmten Positionen kommt. Am Ende jedoch tritt eine Bevorzugung für AG auf.

Bei den letzten beiden Positionen kommt es zu einer Präferenz der Nukleotide AG. Diese Feststellung spiegelt sich im entsprechenden Sequenzlogo des zweiten Exons wieder, welches Abb. 30 wiedergibt. Bis auf die letzten beiden Positionen, also das direkte Ende der Exons, kommt es zu keiner weiteren Bevorzugung, was durch den fehlenden Informationsgehalt an den entsprechenden Positionen ersichtlich wird.

Etwas anders sieht der Fall für das erste Exon aus, hier gibt es eine Bevorzugung für das Nukleotid T an drittletzter Stelle (Daten nicht gezeigt), was jedoch darauf zurückzuführen sein kann, dass bei einigen Genen nur ein Exon annotiert ist, das dann mit einem Stoppcodon endet. Auf die Darstellung des Endes des ersten Exons wird verzichtet. Da bei den Exons 2 bis 4 an drittletzter Position keine Bevorzugung für das Nukleotid T auftrat, deutet dies darauf hin, dass keine direkte Bevorzugung für eine bestimmte Anzahl an Exons in den Genen vorliegt. Hierbei muss ergänzt werden, dass diese Schlussfolgerung nur für die untersuchten Exons 2 bis 4 gilt.

Eine im Vergleich zu den anderen betrachteten Exons verschiedene Nukleotidkomposition tritt bei am Ende des am höchsten nummerierten Exons eines jeden Gens auf, da hier generell am Ende das Stoppcodon lokalisiert ist. Bei einer Darstellung des AT-Gehalts fällt

auf, dass dieser sich zunächst zwischen 50 % und 60 % bewegt (Abb. 31), aber mit abnehmender Distanz zum direkten Endes des Exons deutlich ansteigt. Dies deutet darauf hin, dass die Codon Usage zu AT-reichen Codons hin verlagert wird.

Die nähere Betrachtung der letzten 40 Positionen im höchsten Exon zeigt, dass bis zur Position -4 keine klare Nukleotidbevorzugung zu verzeichnen ist. An der Position -3, das bedeutet 3 Positionen vor dem Ende des Exons, jedoch, tritt eine starke Präferenz für T auf (34, 95 %), sowie vorzugsweise ein A (54,46 %) oder G (40,93 %) an der Position -2 (Daten nicht gezeigt). An der Position -1 tritt vorwiegend das Nukleotid A auf (70,70 %), in etwa 20 % der Exons tritt das Nukleotid G präferenziell auf. Zur näheren Darstellung wurde ein Sequenzlogo erzeugt (Abb. 32). Da jedoch bei Einsatz aller 29993 Sequenzen die Berechnung des Sequenzlogos fehlschlug, wurden stellvertretend nur die ersten 24000 Sequenzen eingesetzt. Bis auf die letzten drei Positionen des höchsten Exons kommt es zu keinerlei Bevorzugung von Nukleotiden bei den eingesetzten Sequenzen, was durch den fehlenden Informationsgehalt deutlich wird. Die letzten drei Positionen jedoch zeigen deutliche Bevorzugungen für T(A/G)(A/G).

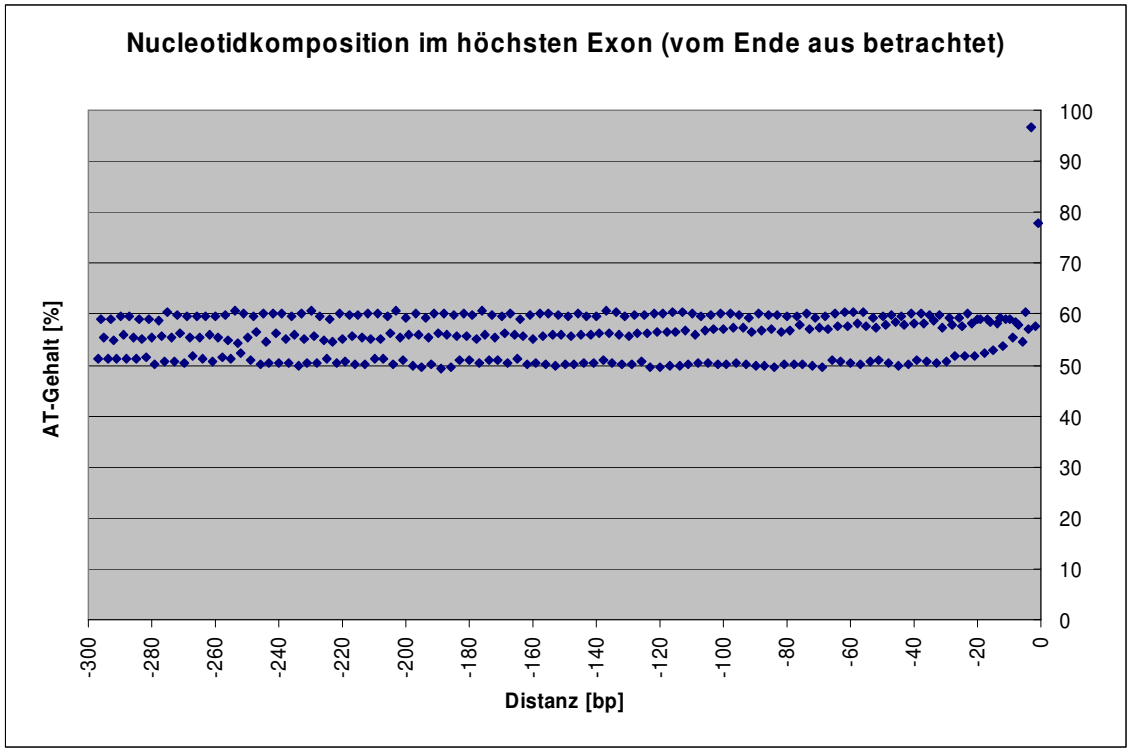


Abb. 31: AT-Gehalt des am höchsten nummerierten Exons der jeweiligen Gene. Die durchschnittliche Länge des höchsten Exons beträgt 157 Nucleotide. Die Sequnzen wurden vom Ende des Exons aus betrachtet und zeigen bis etwa 20 Nukleotide vom Ende der Exons eine Streuung des AT-Gehalts im

Bereich von 50 % und 60 %, wonach der AT-Gehalt ansteigt. Am Ende der Exons tritt der höchste AT-Gehalt auf.

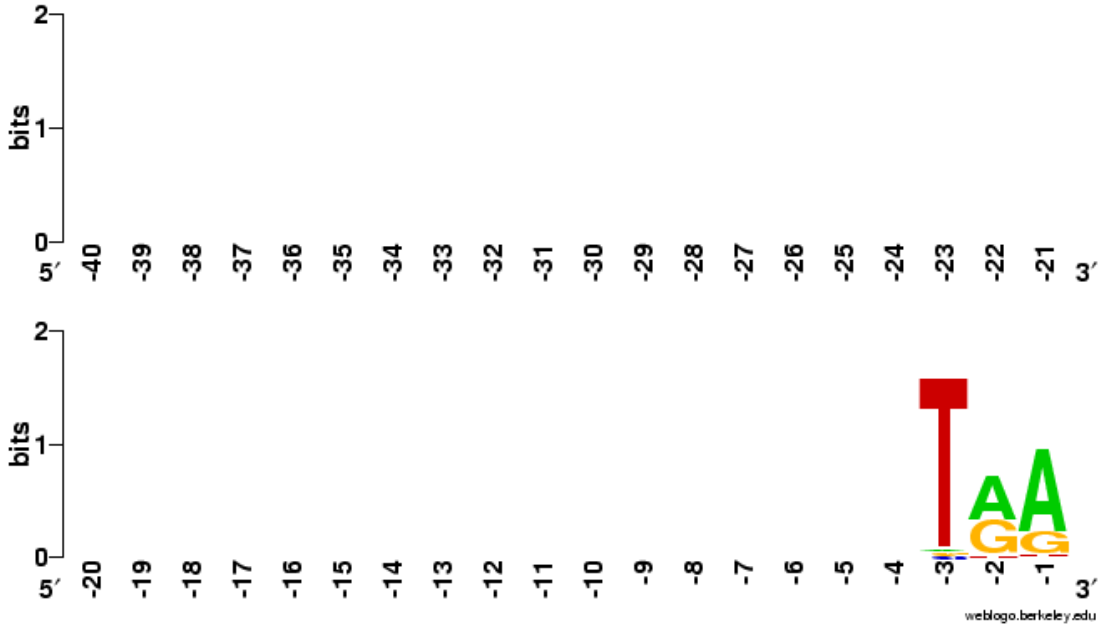


Abb. 32: Sequenzlogo von 24000 der insgesamt 29993 Sequenzen des am höchsten nummerierten Exons der jeweiligen Gene. Betrachtet sind nur die letzten 40 Positionen. Bis auf die letzten drei Positionen ist aufgrund des Informationsgehalts keine weitere Nukleotidpräferenz mehr darstellbar. An den letzten drei Positionen ist das Stoppcodon lokalisiert.

Die erwarteten Stoppcodons in *Arabidopsis thaliana* Genen sind TAA, TAG und TGA. Aufgrund des ungleich häufigen Auftretens der Nucleotide A und G an den letzten beiden Positionen des Sequenzlogos können bevorzugte Stoppcodons abgeleitet werden. Werden die drei Stoppcodons betrachtet, wird an der drittletzten Position zu 100 % ein T erwartet. Unter der Annahme einer gleichmäßigen Bevorzugung aller drei Codons würde an der vorletzten Position zu 66,6 % ein A und zu 33,3 % ein G erwartet werden. Die letzte Position bestünde dann zu 66,6 % aus einem A und zu 33,3 % aus einem G. An der zweitletzten Position im Sequenzlogo (Abb. 32) kommen A und G fast gleich häufig vor, mit einer leichten Präferenz für das Nukleotid A. Das indiziert ein weniger häufiges Auftreten des Stoppcodons T**A**A oder T**A**G sowie ein häufigeres Auftreten des Stoppcodons T**G**A. Beim letzten Nukleotid tritt das erwartete Verhältnis der Nukleotide A und G im Sequenzlogo auf. Dies deutet darauf hin, dass die Stoppcodons TAA und/oder TGA häufiger vorkommen als TAG. Zusammengefasst lässt sich schlussfolgern, dass die beiden Stoppcodons TAA und TGA bevorzugt werden.

3.1.4 3'UTR

Im Folgenden wurde die 3'UTR analysiert, wobei zunächst die erste 3'UTR vom Anfang aus betrachtet wurde. Zu Beginn variiert der AT-Gehalt leicht im Vergleich zu den weiter entfernt liegenden Distanzen (Abb. 33). Generell steigt der AT-Gehalt von etwa 65 % auf knapp unter 70 % an. Die Auftragung der einzelnen Nukleotide in den ersten 40 Positionen zeigt bis auf die generelle Bevorzugung der Nukleotide T und A, was den AT-Gehalt des durchschnittlichen Bereichs der 3'UTR widerspiegelt, keine eindeutige Bevorzugung für Nukleotide an bestimmten Positionen. Daher wurde auf eine Darstellung verzichtet. Auch das entsprechende Sequenzlogo lässt wenige Rückschlüsse auf eine Präferenz zu, jedoch kann festgestellt werden, dass aufgrund des Informationsgehalts eine Präferenz minimal darstellbar ist, wobei an den meisten Positionen eine sehr geringe Bevorzugung für das Nukleotid T vorliegt (Daten nicht gezeigt).

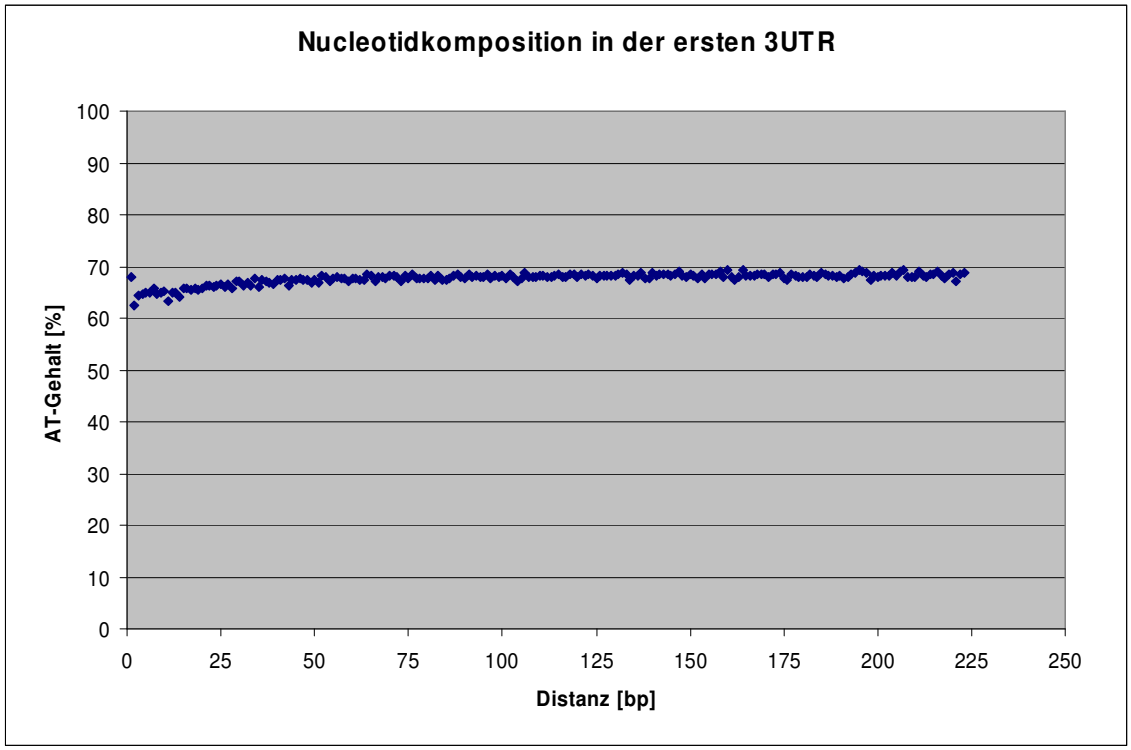


Abb. 33: AT-Gehalt innerhalb der durchschnittlichen Länge (223 bp) der ersten 3'UTR. Zu Beginn ist ein leicht variabler AT-Gehalt zu verzeichnen. Im weiteren Verlauf nähert sich der AT-Gehalt asymptotisch 70 % an.

Anders stellt sich das Bild vom Ende der 3'UTR aus dar, welches anhand der höchsten 3'UTR dargestellt wird (Abb. 34).

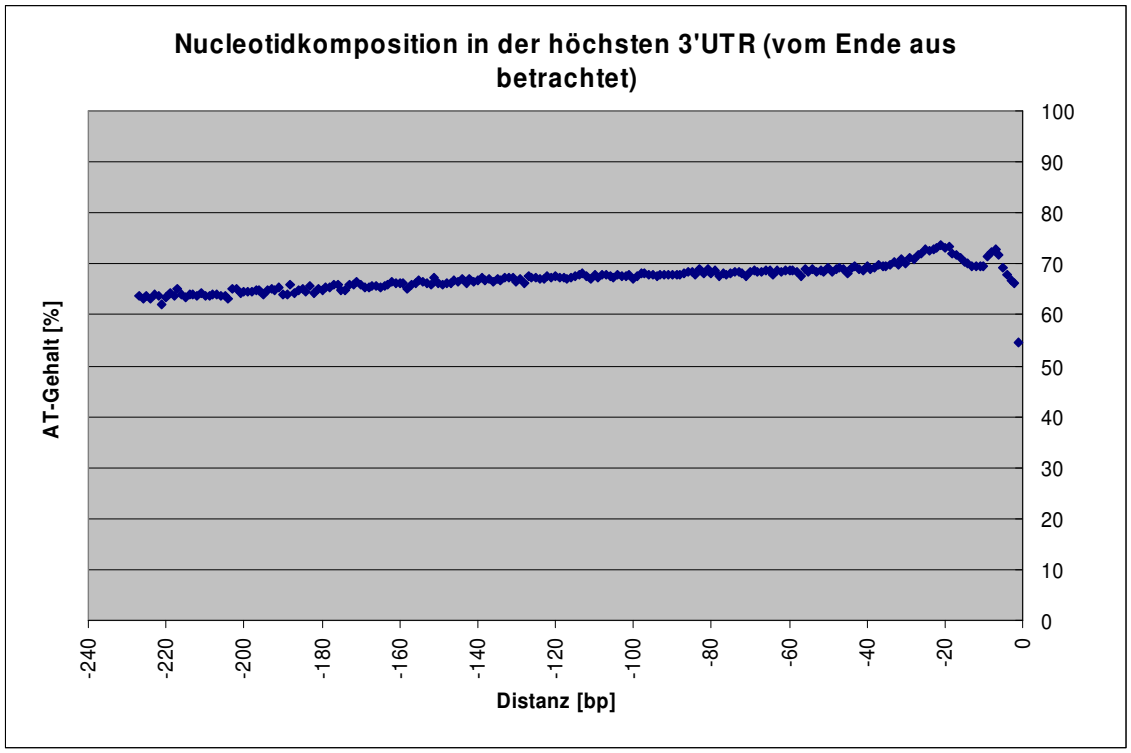


Abb. 34: AT-Gehalt innerhalb von 227 bp der am höchsten nummerierten 3'UTR, welche das Ende der 3'UTR beinhaltet. Der AT-Gehalt steigt mit abnehmender Distanz zum Ende der 3'UTR an, kurz vor dem Ende kommt es zu einem wellenförmigen Verlauf, wobei der AT-Gehalt an der letzten Position, also beim direkten Ende, seinen Minimalwert aufweist.

Der AT-Gehalt steigt bis etwa 30 Positionen vor dem Ende der 3'UTR eher stetig an, wonach ein wellenförmiger AT-Gehalt folgt. Zum Ende hin fällt dieser bis zu seinem Minimalwert ab.

Durch Betrachtung der letzten 40 Positionen kann gezeigt werden (Abb. 35), dass zunächst eine leichte Bevorzugung für das Nukleotid T vorherrscht, welche dann durch A abgelöst wird. Die letzten 12 Positionen sind wieder T-reich.

Des Weiteren ist zu verzeichnen, dass sich die Verteilung der Nukleotide T und A in Abb. 35 gegenläufig zueinander verhält.
Das Sequenzlogo, welches in Abb. 36 dargestellt ist, spiegelt diese Bevorzugung wieder. Der Bereich von -31 bis -25 zeigt die Präferenz für T, wonach sich bis zum Abstand von etwa -14 eine Bevorzugung für A anschließt. Danach tritt wieder eher das Nukleotid T auf.

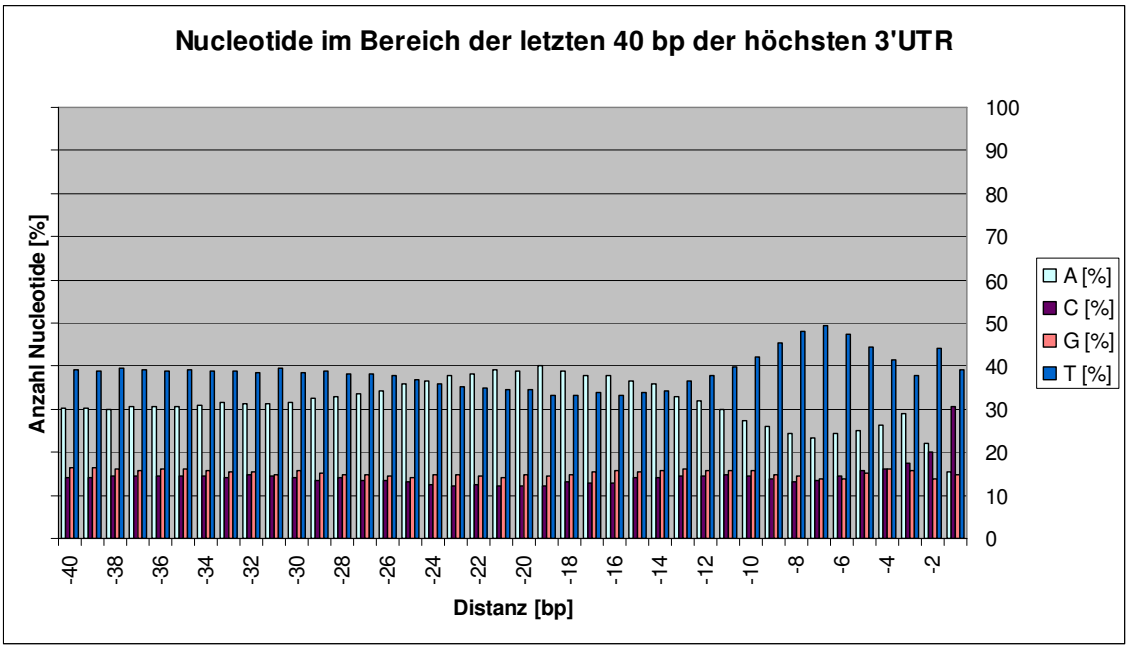


Abb. 35: Betrachtung der Nukleotidkomposition der letzten 40 Positionen der jeweils am höchsten nummerierten 3'UTR, welche diejenige ist, die das Ende der 3'UTR beinhaltet. Generell ist das Ende der 3'UTR AT-reich, wobei sich die Nukleotide A und T gegenläufig zueinander verhalten. Kurz vor dem direkten Ende kommt es zu einem vermehrten Auftreten des Nukleotids T, das Nukleotid C kommt an den letzten beiden Positionen häufiger vor als in den übrigen betrachteten Distanzen.

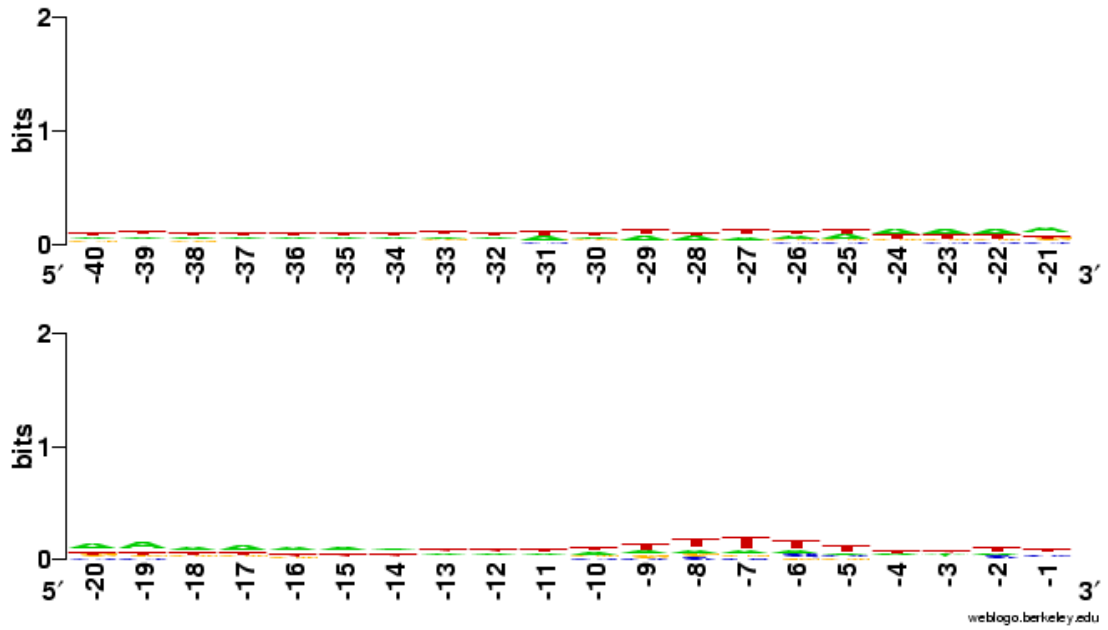


Abb. 36: Sequenzlogo der letzten 40 Positionen der am höchsten nummerierten jeweiligen 3'UTR. Von der Distanz -40 bis -25 tritt das Nukleotid T präferenziell häufiger auf als andere Nukleotide, danach kommt es bis zu einer Distanz von -14 zu einer Bevorzugung von A, wonach es wieder von T abgelöst wird. Im Bereich von -9 bis -5 kommt es zu einem Maximum des Auftretens von T.

3.2 Verteilung von TFBSn relativ zum nächstgelegenen Translationsstart analysiert in 50 bp Schritten

Ausgehend von den Matrizensuchen mit pflanzlichen und nichtpflanzlichen Matrizen (Kapitel 2.2.1 und 2.2.2), welche mit Patser (Kapitel 2.4) durchgeführt wurden, erfolgte der Import der putativen Bindungsstellen der einzelnen Matrizen unter Verwendung der TFBSitesTools in die Datenbank (Kapitel 2.5.3). Anschließend wurde eine Redundanzanalyse für die putativen TFBSn jeder Matrizensuche durchgeführt, gefolgt von dem Einlesen des nächstgelegenen Gens anhand des annotierten Translationsstarts dieses Gens inklusive der relativen Distanz zu diesem Translationsstartpunkt. Wies das Gen einen annotierten Transkriptionsstart auf, wurde die relative Distanz hierzu ebenfalls in die Datenbank aufgenommen. Analog zu der Vorgehensweise mit den putativen TFBSn der Matrizen wurde mit den erzeugten 42 x 7500 Zufallspositonen verfahren. Es folgte eine Verteilungsanalyse mit den putativen TFBSn jeder Matrize bzw. jedem Satz an Zufallspositionen relativ zum Translationsstart mit den TFBSitesTools.

Bei dieser Analyse wurden die Distanzen in upstream und downstream eingeteilt, wobei upstream bedeutet, dass die putative Bindungsstelle (oder generell Position) 5' vom zugeordneten Translationsstartpunkt liegt, und 3', dass diese Bindungsstelle 3' vom Bezugspunkt auftritt. Da diese Analyse eine Einschätzung über die prinzipielle Verteilung von putativen Transkriptionsfaktor-Bindungsstellen liefern sollte, wurde das Auftreten dieser in 50 bp Schritten zusammengefasst und grafisch dargestellt.

Zunächst wird auf die Verteilung der Zufallspositionen eingegangen. Diese Positionen vermitteln einen Eindruck über eine zufällige Verteilung. Dadurch kann anschließend ein Vergleich mit den Verteilungen von matrizenbasierenden putativen Bindungsstellen durchgeführt werden.

3.2.1 Verteilung von Zufallspositionen relativ zum nächstgelegenen Translationsstart

Es wurden 42 x 7500 Zufallspositionen erzeugt, mit denen wie in Kapitel 3.2 beschrieben verfahren wurde. Anschließend wurden Verteilungsanalysen relativ zum nächstgelegenen Translationsstartpunkt durchgeführt, wobei das Auftreten von Positionen in 50 bp Schritten zusammengefasst und dargestellt wurde. Positionen (oder putative Transkriptionsfaktor-Bindungsstellen) können entweder vor einem Gen, also upstream, lokalisiert sein, oder nach dem Start eines Gens, downstream, liegen. Ein Nukleotidbereich von -10000 bp bis +10000 bp wurde bei den Analysen untersucht. Beispielhaft ist in Abb. 37 eine typische Verteilung der Positionen dargestellt. Die mit den anderen 41 Zufallsdatensätzen ermittelten

Verteilungen zeigten gleiche Verteilungsmuster. Generell lässt sich feststellen, dass mehr Positionen im downstream-Bereich auftreten als dies im upstream-Bereich der Fall ist, wobei das Maximum im downstream-Bereich aufzufinden ist.

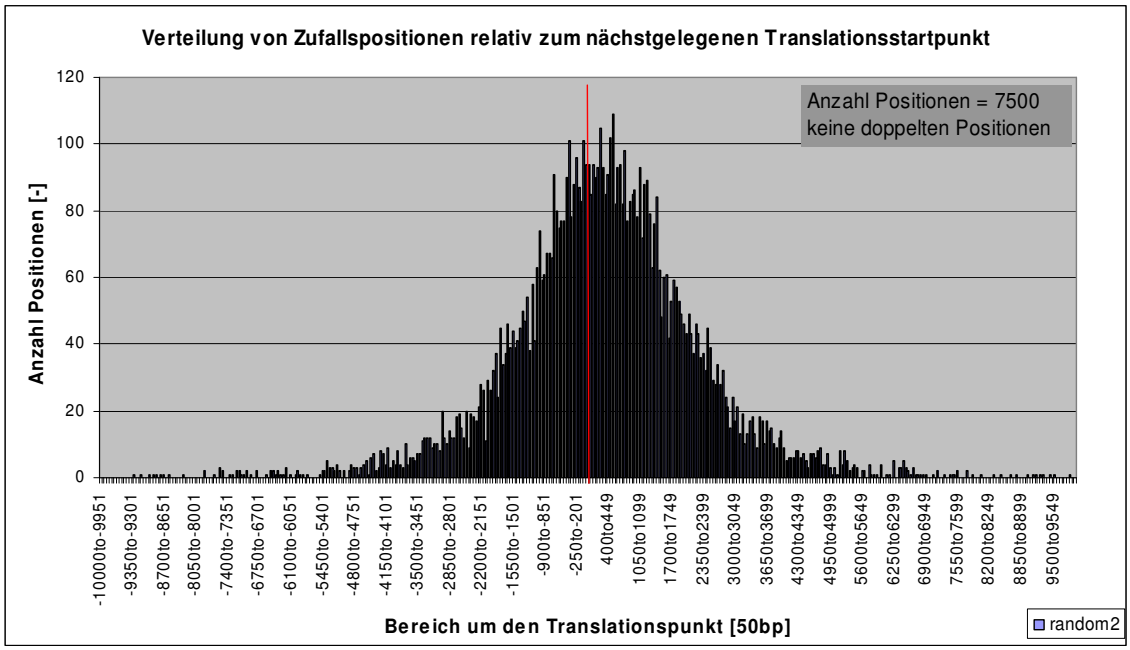


Abb. 37: Typische Verteilung von Zufallspositionen relativ zum nächstgelegenen Translationsstartpunkt aufgeteilt in 50 bp Schritte. Der senkrechte rote Strich deutet auf den Translationsstartpunkt hin. Im upstream-Bereich sind weniger Positionen lokalisiert als im downstream-Bereich, wo auch das Maximum vorzufinden ist.

Mit den 42 Datensätzen wurden unter Verwendung von STATeasy (Kapitel 2.6.2) der U-Test sowie der Median-Test eingesetzt, wodurch bestätigt werden konnte, dass alle die gleiche Verteilungsform aufweisen. Dies deutet darauf hin, dass die Zufallspositionen tatsächlich zufällig generiert wurden und kein systematischer Fehler vorliegt. Optisch weisen die Werte jeweils Glockenkurven auf, was auf eine Normalverteilung der Werte hindeutet. Die Normalverteilung wurde mittels des Kolmogoroff-Smirnoff Anpassungstests (modifiziert nach Kuiper) sowie des Shapiro-Wilk Anpassungstests überprüft (Kapitel 2.6.2), jedoch konnte die Normalverteilung nicht nachgewiesen werden. Das kann auf einen unbekannten biologischen Einfluss auf die Positionen zurückzuführen sein, oder aber aus einer formalen Signifikanz resultieren, die aufgrund einer großen Stichprobenanzahl auftreten kann (http://www.lrz-muenchen.de/services/schulung/unterlagen/grundbegriffe/grundbegriffe-10.html). Auf eine Stichprobenverkleinerung durch Zusammenfassung der einzelnen empirischen Werte wurde verzichtet.

Ziel im Folgenden war es nun, einen direkten Vergleich der Verteilung der Zufallspositionen mit den matrizenbasierenden Transkriptionsfaktor-Bindungsstellen zu ermöglichen. Zu diesem Zweck wurde bei den Zufallspositionen die Anzahl der Positionen im upstream- und downstream-Bereich getrennt aufsummiert und der prozentuale Anteil in den beiden Bereichen bestimmt. Neben dem Mittelwert für die prozentuale Anzahl der Positionen in den entsprechenden Bereichen wurden ebenfalls die beiden Maximalwerte erfasst (Tabelle 6).

Tabelle 6: Durchschnittlicher Anteil der Zufallspositionen im upstream-Bereich sowie im downstream-Bereich relativ zum nächstgelegenen Translationsstartpunkt, angegeben in Prozent. Ausgang sind die 42 x 7500 erzeugten Positionen. Die minimale sowie maximale Anzahl an Positionen in den beiden Bereichen sind ebenfalls aufgeführt.

Ø Anteil Positionen upstream [%]	41,68	Ø Anteil Positionen downstream [%]	58,32
min. Wert upstream [%]	40,69	min. Wert downstream [%]	57,10
max. Wert upstream [%]	42,90	max. Wert downstream [%]	59,31

Basierend auf der Verteilung der Zufallspositionen wurden für die Auswertung der Verteilungen der putativen Transkriptionsfaktor-Bindungsstellen die folgenden Kriterien definiert.

Eine indifferente Verteilung der putativen TFBSn lag vor, wenn die Abweichung der TFBSn einseitig nicht mehr als 10 % über dem Durchschnittswert für die Zufallspositionen lag.
Eine upstreamlastige Verteilung hingegen lag vor, wenn diese nicht indifferent war, jedoch die upstream-Abweichung mehr als 10 % vom Durchschnittswert für die Zufallspositionen betrug.
Eine Verteilung war downstreamlastig, wenn sie nicht indifferent war sowie die downstream-Abweichung der TFBSn mehr als 10 % über dem Durchschnittswert für die Zufallspositionen lag.

Im Folgenden werden nun die Verteilungsanalysen der putativen Transkriptionsfaktor-Bindungsstellen relativ zum nächstgelegenen Translationsstartpunkt dargestellt und bewertet.

3.2.2 Verteilung von putativen pflanzlichen TFBSn relativ zum nächstgelegenen Translationsstart

Mit den putativen pflanzlichen Transkriptionsfaktor-Bindungsstellen der 42 Matrizen sowie mit den putativen Bindungsstellen der entsprechenden Zufallsmatrizen für die pflanzlichen Matrizen wurde, wie in Kapitel 3.2 beschrieben, verfahren. Durch eine grafische Auswertung konnten drei prinzipielle Verteilungsarten festgestellt werden. Es traten Anreicherungen der TFBSn eher im upstream-Bereich, eine ähnliche Verteilung wie die der Zufallspositionen (indifferent), sowie eine vermehrte Anreichung von TFBSn im downstream-Bereich auf.

Es folgte mittels der in Kapitel 3.2.1 definierten Kriterien die Einteilung in die drei Verteilungsformen upstreamlastig, indifferent und downstreamlastig durch Vergleich der prozentualen Verteilung mit denen der Zufallspositionen. In Tabelle 7 ist die Anzahl der pflanzlichen Matrizen und der entsprechenden Zufallsmatrizen bezüglich der jeweiligen Verteilungstypen aufgeführt.

Tabelle 7: Angeführt sind die nach Anwendung des Kriteriums aus Kapitel 3.2.1 identifizierten Verteilungstypen der pflanzlichen Matrizen und ihrer entsprechenden Zufallsmatrizen.

Verteilungstyp	**Anzahl pflanzliche Matrizen (%)**	**Anzahl Zufallsmatrizen der pflanzlichen Matrizen (%)**
upstreamlastig	9 (21,43 %)	6 (14,29 %)
downstreamlastig	15 (35,71 %)	30 (71,43 %)
indifferent	18 (42,86 %)	6 (14,29 %)

Unter Verwendung der Kriterien konnte gezeigt werden (Tabelle 7), dass die pflanzlichen Matrizen eine leichte Präferenz für eine indifferente Verteilung (18 Matrizen) sowie für eine downstreamlastige Verteilung aufwiesen (15 Matrizen). Die upstreamlastige Verteilung kam hingegen weniger häufig vor (9 Matrizen). Die entsprechenden Zufallsmatrizen für die pflanzlichen Matrizen wiesen eine Bevorzugung für die downstreamlastige Verteilung (30 Matrizen) auf, die upstreamlastige und die indifferente Verteilung traten jeweils gleich häufig, aber weniger häufig als die downstreamlastige Verteilung auf. Beispielhaft sind die drei Verteilungsformen in Abb. 38 bis Abb. 40 für pflanzliche Matrizen dargestellt.

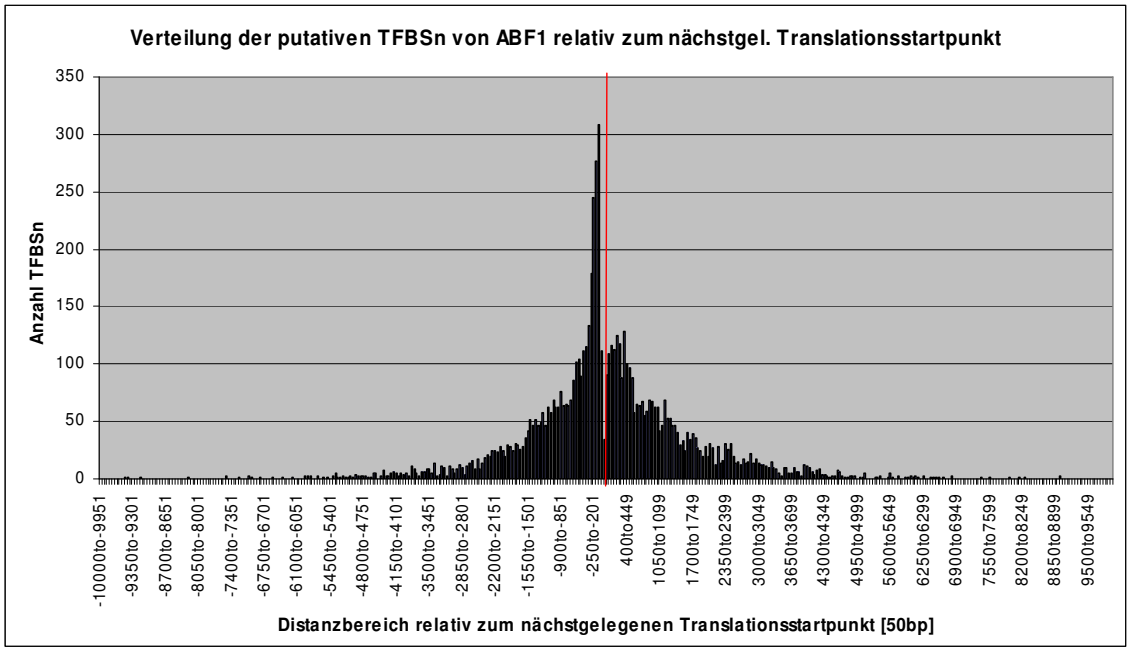


Abb. 38: Beispiel für eine upstreamlastige Verteilungsform. Dargestellt ist diese für die Transkriptionsfaktor-Bindungsstellen des pflanzlichen Faktors ABF1. Die Distanz relativ zum nächstgelegenen Translationsstartpunkt ist in 50 bp- Schritten zusammengefasst.

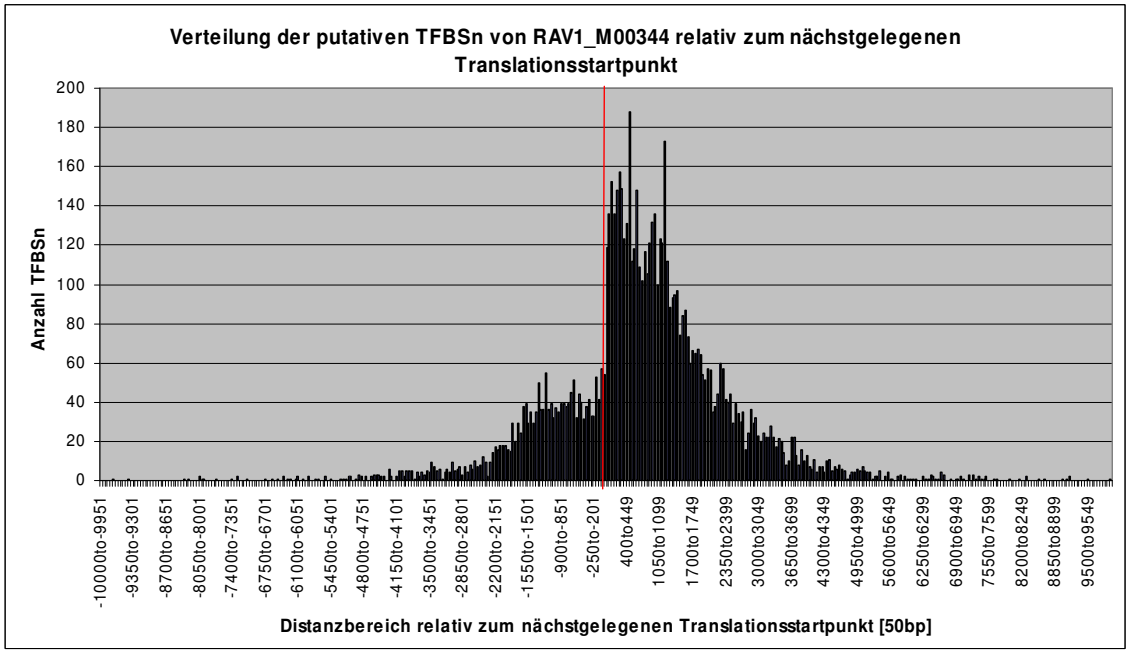


Abb. 39: Beispiel für eine downstreamlastige Verteilungsform. Dargestellt ist diese für die Transkriptionsfaktor-Bindungsstellen des pflanzlichen Faktors RAV1. Die Distanz relativ zum nächstgelegenen Translationsstartpunkt ist in 50 bp- Schritten zusammengefasst.

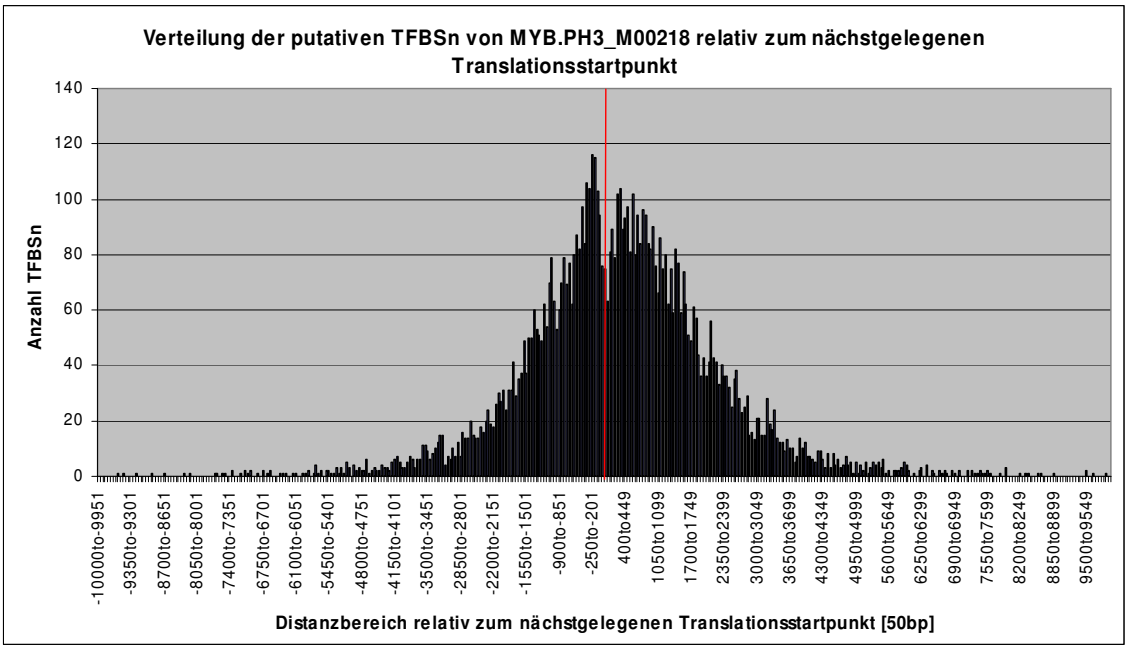


Abb. 40: Beispiel für eine indifferente Verteilungsform. Dargestellt ist diese für die Transkriptionsfaktor-Bindungsstellen des pflanzlichen Faktors MYB.PH3. Die Distanz relativ zum nächstgelegenen Translationsstartpunkt ist in 50 bp- Schritten zusammengefasst.

Um den Einfluss des AT-Gehaltes auf die Position der putativen TFBSn zu analysieren, wurde die Verteilungsform jeder pflanzlichen Matrize mit ihrer Zufallsmatrize direkt miteinander verglichen. Hierbei zeigte sich, dass bei 25 von 42 pflanzlichen Matrizen (60 %) eine gleiche Art der Verteilung vorlag.

Die abweichenden Verteilungen der pflanzlichen Matrizen von ihrer jeweiligen Zufallsmatrizen werden in Tabelle 8 zusammengefasst. Diese trat bei 17 Matrizen auf. Von diesen weisen die meisten eine Veränderung der Verteilung von der indifferenten Verteilung der pflanzlichen Matrize hin zu einer downstreamlastigen Verteilung der entsprechenden Zufallsmatrize auf (in 12 Fällen). Bei drei Matrizen kam es zu einer starken Änderung der Verteilung von der upstreamlastigen zur downstreamlastigen Verteilungsform.

Eine abweichende Verteilungsform tritt vorwiegend bei Transkriptionsfaktoren der Klasse bZIP auf, zu denen bZIP910, bZIP911, TGA1, ABF1, O2 und TGA1a gehören (dies sind 7 der 8 verwendeten Matrizen). bZIP Faktoren erkennen vorwiegend die Kernsequenz ACGT. Eine weitere Klasse, die diese Sequenz erkennt, ist bHLH. In diesen Analysen wurden zwei bHLH Matrizen (PIF3) eingesetzt. Diese beiden Matrizen zeigten ebenfalls eine Verteilungsänderung. Eine weitere Auffälligkeit wird bei der Klasse TCP, die durch die beiden Faktoren PCF2 und PCF5 vertreten ist, entdeckt. Hier wird ebenfalls bei allen eingesetzten Matrizen eine Änderung der Verteilungsform festgestellt.

Zusammengefasst deutet die prinzipielle Änderung der Verteilungsform zwischen pflanzlichen Matrizen und ihren entsprechenden Zufallsmatrizen darauf hin, dass die Nukleotidkomposition nicht der alleinige Parameter für die Art der Verteilung sein kann. Bei einigen Mitgliedern von Transkriptionsfaktor-Klassen kommt es gehäuft zu dieser Änderung der Verteilungsform.

Tabelle 8: Auflistung der Namen der Matrizen für pflanzliche Transkriptionsfaktoren, deren dazugehörige Zufallsmatrize eine unterschiedliche Art der Verteilung aufweist. Vorwiegend kommt es zu einer Änderung von der indifferenten Verteilung hin zu einer downstreamlastigen Verteilung. Angegeben ist ebenfalls die prozentuale Abweichung von den Zufallspositionen.

Matrize	**Verteilungstyp der pflanzlichen Matrize**	**Abweichung [%]**	**AT-Gehalt [%]**	**Verteilungstyp der Zufalls-matrize der pflanzl. Matrize**	**AT-Gehalt [%]**	**Abweichung [%]**
AG	indifferent	5,76	63,8	down	64,39	-20,37
AGP1	indifferent	-8,85	61,05	down	59,47	-23,86
ATHB5	indifferent	4,12	61,19	down	63,71	-21,5
bZIP910(1)	indifferent	-5,17	36,57	down	37,04	-28,26
bZIP911(1)	indifferent	-10,33	31,06	down	28,41	-34,11
bZIP911(2)	indifferent	-7,07	45,00	down	45,83	-26,22
MYB.PH3(2)	indifferent	6,87	70,73	up	69,23	22,73
PCF2	indifferent	5,42	26,01	down	23,95	-29,83
PCF5	indifferent	-10,25	23,77	down	26,27	-31,96
PIF3(1)	indifferent	3,49	38,07	down	36,83	-35,79
PIF3(2)	indifferent	3,72	36,03	down	35,66	-30,82
TGA1	indifferent	-6,2	52,69	down	48,92	-22,93
ZAP1	indifferent	-7,99	44,69	down	42,47	-47,05
ABF1	up	28,93	35,56	down	39,44	-28,67
ID1	up	19,74	66,06	indifferent	67,27	3,19
O2	up	26,26	40,91	down	40	-16,91
TGA1a	up	15,73	37,27	down	40	-40,98

Analog zu der Verfahrensweise mit den pflanzlichen Matrizen sowie ihren entsprechenden Zufallsmatrizen wurde mit den nichtpflanzlichen verfahren, deren Ergebnisse im nächsten Abschnitt dargestellt werden.

3.2.3 Verteilung von putativen nichtpflanzlichen TFBSn relativ zum nächstgelegenen Translationsstart

Mit den putativen nichtpflanzlichen Transkriptionsfaktor-Bindungsstellen der 27 Matrizen sowie mit den putativen Bindungsstellen der entsprechenden Zufallsmatrizen für die nichtpflanzlichen Matrizen wurde wie in Kapitel 3.2 beschrieben verfahren.

Es folgte analog zu der Vorgehensweise bezüglich der pflanzlichen Matrizen mittels der in Kapitel 3.2.1 definierten Kriterien die Einteilung in die drei Verteilungsformen upstreamlastig, indifferent und downstreamlastig durch Berücksichtigung der prozentualen Abweichung von den Zufallspositionen. In Tabelle 9 ist die Anzahl der nichtpflanzlichen Matrizen und der Zufallsmatrizen für die nichtpflanzlichen Matrizen bezüglich ihrer Verteilung aufgeführt.

Tabelle 9: Angeführt sind die nach Anwendung des Kriteriums aus Kapitel 3.2.1 identifizierten Verteilungstypen der nichtpflanzlichen Matrizen sowie ihrer entsprechenden Zufallsmatrizen.

Verteilungstyp	Anzahl nichtpflanzliche Matrizen (%)	Anzahl Zufallsmatrizen der nichtpflanzlichen Matrizen (%)
upstreamlastig	1 (3,70 %)	2 (7,41 %)
downstreamlastig	16 (59,26 %)	23 (85,19 %)
indifferent	10 (37,04 %)	2 (7,41 %)

Im Gegensatz zu der Auswertung bei den pflanzlichen Matrizen ist bei den nichtpflanzlichen Matrizen eine vorwiegende downstreamlastige Verteilung zu verzeichnen. 10 Matrizen zeigten eine indifferente Verteilung, während die upstreamlastige Verteilung die Ausnahme war und nur in einem Fall auftrat. Überraschend ist, dass auch die nichtpflanzlichen Matrizen häufig keine indifferente Verteilung aufwiesen. Die entsprechenden Zufallsmatrizen für die nichtpflanzlichen Matrizen weisen eine etwas andere Aufteilung in die drei Verteilungstypen auf. Hier ist wie bei den Zufallsmatrizen der pflanzlichen Matrizen eine downstreamlastige Verteilungsform die am häufigsten auftretende Form. Upstreamlastige und indifferente Verteilungen sind eher die Ausnahme (jeweils 2 Matrizen). Trotz der eingesetzten aus Bindungsstellen bestehenden Matrizen, die jedoch Bindungsstellen in nichtpflanzlichen Systemen beschreiben, unterscheiden sie sich von den reinen Zufallsmatrizen. Sie unterscheiden sich durch den vorgegebenen AT-Gehalt und einer durch die jeweilige Matrize beschriebenen DNA-Sequenz.

Es folgte ein direkter Vergleich der Art der Verteilung der nichtpflanzlichen Matrize mit ihrer dazugehörigen Zufallsmatrize, wobei festgestellt werden konnte, dass es in 63 % der Fälle (17 Matrizen) zu einer gleichen Verteilungsform kam. Die 10 nichtpflanzlichen Matrizen, deren dazugehörige putative Transkriptionsfaktor-Bindungsstellen sich in ihrer

Verteilungsform von ihrer entsprechenden Zufallsmatrize unterscheiden, sind in Tabelle 10 aufgelistet. Es kommt vorwiegend zu einer Änderung der Verteilung von indifferent zu downstreamlastig.

Auch diese Ergebnisse deuten darauf hin, dass der AT-Gehalt nicht der alleinige Parameter für die Art der Verteilung relativ zum nächstgelegenen Translationsstartpunkt sein kann.

Tabelle 10: Auflistung der Namen der Matrizen für nichtpflanzliche Transkriptionsfaktoren, deren dazugehörige Zufallsmatrize eine unterschiedliche Art der Verteilung aufweist. Vorwiegend kommt es zu einer Änderung von der indifferenten Verteilung hin zu einer downstreamlastigen Verteilung. Angegeben ist ebenfalls die Abweichung von den Zufallspositionen.

Matrize	**Verteilungs-typ der nicht-pflanzliche Matrize**	**Abweichung [%]**	**AT-Gehalt [%]**	**Verteilungs-typ der Zufallsmatrize der nichtpflanzl. Matrize**	**AT-Gehalt [%]**	**Abweichung [%]**
AREB6_M00415	down	-24,2	54,63	indifferent	59,26	-13,51
COUP_M00158	indifferent	-12,26	52,2	down	52,2	-34,51
PPAR-alpha_M00242	indifferent	-12,88	54,29	down	53,57	-31,26
E2F_M00024	indifferent	8,55	50,67	down	56,00	-30,79
HAP3_M00288	indifferent	-6,07	52,23	down	51,79	-24,25
Pax-3_M00360	indifferent	-11,42	47,63	down	53,25	-42,97
Pax-6_M00097	indifferent	-12,95	57,53	down	57,24	-40,61
POU2F1_M00136	indifferent	0,34	63,7	up	66,67	18,04
Dl_M00120	indifferent	-0,62	56,06	down	56,06	-27,03
Brachyury_M00150	indifferent	-7,74	55,62	down	55,83	-36,39

3.3 Vorkommen von putativen TFBSn in den einzelnen Bereichen der Gene und im intergenischen Bereich

Ausgehend von den erzeugten Zufallspositionen sowie den mit den Matrizen bestimmten putativen Transkriptionsfaktor-Bindungsstellen wurde die Lokalisierung jeder einzelnen Position mit den TFBSitesTools (Kapitel 2.5.3) identifiziert und die relative Distanz zum Startpunkt jedes Bereichs protokolliert. Zunächst wurde das einfache Vorhandensein der Bindungsstellen in diesen Bereichen erfasst („Einfache Lokalisierung von Transkriptionsfaktor-Bindungsstellen in den bestimmten Bereichen innerhalb der Gene und im intergenischen Bereich“ in Kapitel 2.5.3). Für die Zufallspositionen wurden diese Werte ebenfalls ermittelt, um einen Vergleichswert oder auch Erwartungswert zu erhalten, unter der Voraussetzung, dass die Transkriptionsfaktor-Bindungsstellen zufällig verteilt sind. In weiteren Analysen wurden die einzelnen Bereiche innerhalb der Gene detaillierter betrachtet.

3.3.1 Lokalisierung von putativen TFBSn im Intron, Exon, den UTRs und im intergenischen Bereich

Zunächst wurde die einfache Lokalisierung der Zufallspositionen und der putativen Transkriptionsfaktor-Bindungsstellen in den jeweiligen Bereichen betrachtet. Da die Werte aufgrund der unterschiedlichen Anzahl der detektierten TFBSn bzw. Positionen untereinander nicht vergleichbar sind, wurden diese in Prozente umgerechnet. Im Fall der Zufallspositionen wurde aus den 42 x 7500 einzelnen Positionen, bezogen auf die aufgetretenen Positionen in den einzelnen Bereichen, zuvor ein Gesamtmittelwert gebildet. Die prozentuale Anzahl an putativen Transkriptionsfaktor-Bindungsstellen der pflanzlichen Matrizen, der Zufallsmatrizen der pflanzlichen Matrizen, der nichtpflanzlichen Matrizen sowie der Zufallsmatrizen der nichtpflanzlichen Matrizen im intergenischen Bereich, in der 5’UTR, in Introns, in Exons und in der 3’UTR sind in Abb. 41 aufgeführt. Zum Vergleich sind die Daten für die Zufallspositionen ebenfalls dargestellt. Im intergenischen Bereich sind etwa 44 % der Zufallspositionen lokalisiert, die restlichen 56 % liegen innerhalb von Genen. Diese Werte stimmen gut mit der Nukleotidaufteilung des Genoms überein, denn 43,7 % der Nukleotide befinden sich im intergenischen Bereich und 56,3 % innerhalb von Genen (Kapitel 3.1). Innerhalb von Genen kommen die meisten Zufallspositionen in Exons vor, gefolgt von Introns, der 3’UTR sowie der 5’UTR (Abb. 41). Auch dieses Ergebnis stimmt mit der Aufteilung der Nukleotide auf die prinzipiellen Bereiche innerhalb von Genen überein (Kapitel 3.1).

Im Vergleich zu der prozentualen Anzahl an Zufallspositionen sind weniger putative Bindungsstellen der Matrizen im intergenischen Bereich lokalisiert. Die putativen

Bindungsstellen der pflanzlichen Matrizen weisen den zweithöchsten Wert auf, die der Zufallsmatrizen der pflanzlichen Matrizen kommen weniger häufig vor, jedoch häufiger als die der nichtpflanzlichen Matrizen. Die wenigsten Bindungsstellen treten bei den Zufallsmatrizen der pflanzlichen Matrizen auf. In der 5'UTR weisen die Zufallspositionen sowie die putativen Transkriptionsfaktor-Bindungsstellen etwa gleich hohe Werte auf, wobei diese sehr niedrig sind. In Introns zeigen die Zufallspositionen den höchsten Wert. Die Bindungsstellen der pflanzlichen sowie nichtpflanzlichen Matrizen kommen in etwa gleich häufig vor. Die Bindungsstellen der Zufallsmatrizen weisen jeweils ebenfalls etwa ähnliche Werte auf, diese sind jedoch niedriger als die für ihre pflanzlichen bzw. nichtpflanzlichen Matrizen. In Exons ist die prozentuale Anzahl an putativen Transkriptionsfaktor-Bindungsstellen größer als die der Zufallspositionen. Der höchste Wert ist für die Zufallsmatrizen der nichtpflanzlichen Matrizen zu verzeichnen, gefolgt von den nichtpflanzlichen Matrizen, den Zufallsmatrizen der pflanzlichen Matrizen sowie den pflanzlichen Matrizen. In der 3'UTR ist die prozentuale Anzahl an Zufallspositionen sowie Bindungsstellen der Matrizen recht gering und die Werte liegen etwa auf einem Niveau (Abb. 41).

Im Folgenden wurde die Analyse auf die für spätere Analysen festgelegten Genbereiche erweitert (Kapitel 2.1.1). Es wurden also die erste 5'UTR, die ersten 5 Introns, die ersten 4 Exons sowie die erste 3'UTR näher betrachtet.

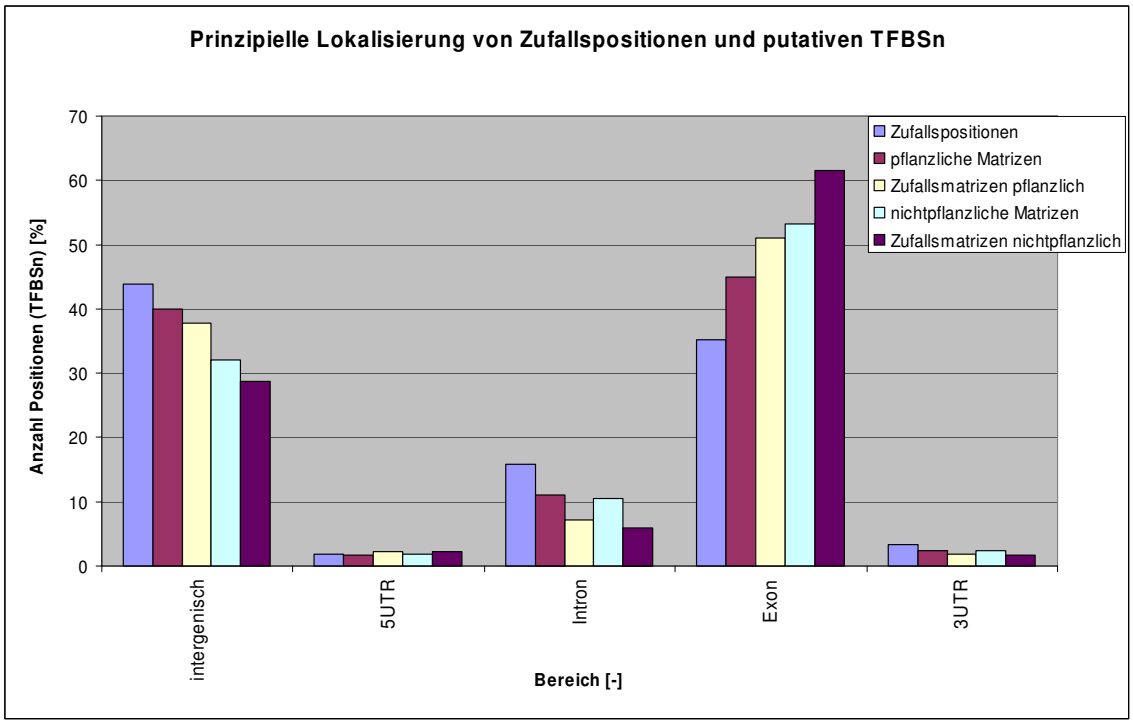


Abb. 41: Prozentuale Anzahl der Zufallspositionen bzw. der putativen Transkriptionsfaktor-Bindungsstellen im intergenischen Bereich, in der 5'UTR, in Introns, Exons und der 3'UTR. Die exakten Werte befinden sich im Anhang (Tabelle 41).

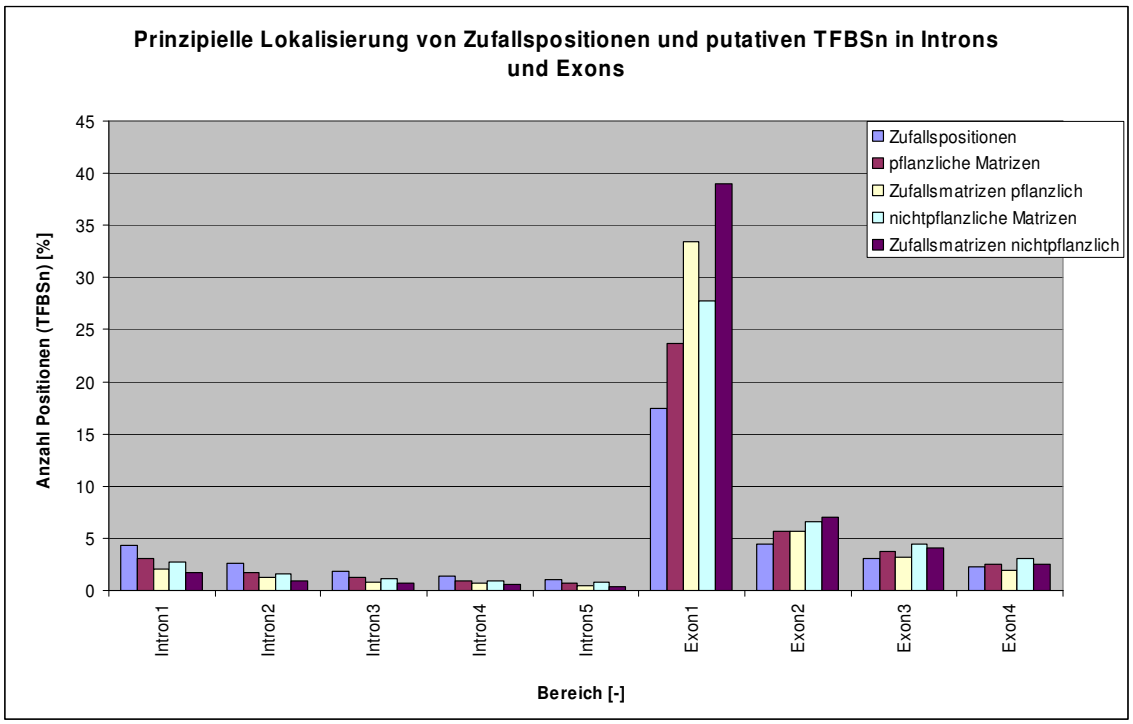


Abb. 42: Prozentuale Anzahl der Zufallspositionen bzw. der putativen Transkriptionsfaktor-Bindungsstellen in den ersten fünf Introns sowie den ersten vier Exons. Die exakten Werte sind im Anhang aufgeführt (Tabelle 42).

Die Werte in der ersten 5'UTR und der ersten 3'UTR unterscheiden sich nur im Nachkommastellenbereich von den Werten für die gesamte 5'UTR und die gesamte 3'UTR, die in Abb. 41 dargestellt sind. Daher wurde auf eine nähere Darstellung der ersten 5'UTR und der ersten 3'UTR verzichtet. Die prozentuale Anzahl der Zufallspositionen sowie der mit Matrizen ermittelten putativen Transkriptionsfaktor-Bindungsstellen in den ersten fünf Introns und in den ersten vier Exons ist in Abb. 42 angeführt. Die Feststellung, dass mehr Zufallspositionen im Intron lokalisiert sind als dies für die putativen TFBSn der Matrizen zutrifft, kann auch für die betrachteten ersten fünf Introns bestätigt werden. Wie bei der Analyse der Introns werden in den ersten fünf Introns mehr putative Bindungsstellen, detektiert mit pflanzlichen und nichtpflanzlichen Matrizen, aufgefunden als dies bei ihren entsprechenden Zufallsmatrizen der Fall ist. Generell gesehen nimmt die Anzahl der Zufallspositionen oder auch putativen Bindungsstellen mit zunehmendem Intron ab, was vermutlich damit zusammenhängt, dass die Anzahl der Gene, die diese Introns aufweisen, mit zunehmender Intronanzahl abnimmt. Die meisten Zufallspositionen und putativen TFBSn in Abb. 41 treten im ersten Exon auf, wobei die Zufallspositionen deutlich weniger Positionen in dem Bereich aufweisen als bei den putativen Bindungsstellen beobachtet. Bei den matrizenbasierenden putativen Transkriptionsfaktor-Bindungsstellen treten die der Zufallsmatrizen häufiger auf als die der pflanzlichen bzw. nichtpflanzlichen Matrizen. Der Grund für das vermehrte Auftreten von Zufallspositionen bzw. matrizenbasierten putativen Transkriptionsfaktor-Bindungsstellen im ersten Exon im Vergleich zu den anderen drei Exons dürfte weitestgehend an der durchschnittlichen Länge des ersten Exons liegen, die 692 bp beträgt, während das zweite Exon nur noch eine durchschnittliche Länge von 259 bp aufweist. Beim zweiten Exon ist die prozentuale Anzahl an Zufallspositionen im Vergleich zu den putativen Bindungsstellen geringer. Wenig größer sind die Werte für die pflanzlichen Matrizen sowie deren Zufallsmatrizen. Die Bindungsstellen der nichtpflanzlichen Matrizen und die der Zufallsmatrizen der nichtpflanzlichen Matrizen weisen leicht höhere Werte als der pflanzlichen Matrizen auf. Die prozentuale Anzahl an Zufallspositionen in den Exons 3 und 4 ist ebenfalls größer als für die matrizenbasierenden Bindungsstellen. Jedoch liegen die Werte zunehmend dichter beieinander. Während die Zufallsmatrizen in diesen Exons ähnlich hohe Werte aufweisen wie die Zufallspositionen, und die Werte der Zufallsmatrizen der nichtpflanzlichen Matrizen nur leicht höher sind, treten im Vergleich mehr putative Bindungsstellen für die pflanzlichen und nichtpflanzlichen Matrizen auf.

Nachdem auf die prinzipielle Lokalisierung der Zufallspositionen und der mit Matrizen detektierten putativen Transkriptionsfaktor-Bindungsstellen der pflanzlichen Matrizen, der Zufallsmatrizen für die pflanzlichen Matrizen sowie der nichtpflanzlichen Matrizen und deren

Zufallsmatrizen dargestellt wurde, erfolgt nun der direkte Vergleich jeder einzelnen Matrize im Vergleich zu den Zufallspositionen.

3.3.2 Lokalisierung von putativen TFBSn in den einzelnen Bereichen der Gene

Bislang wurde die Lokalisierung der Summe der Transkriptionsfaktor-Bindungsstellen aller pflanzlichen Matrizen, aller Zufallsmatrizen für die pflanzlichen Matrizen, der nichtpflanzlichen Matrizen sowie der Zufallsmatrizen für die nichtpflanzlichen Matrizen im Vergleich zu den Zufallspositionen im intergenischen Bereich und innerhalb von Genen dargestellt. Die putativen Bindungsstellen einer einzelnen Matrize gingen dabei zwar in die Summe ein, wurden aber bislang nicht gesondert betrachtet. Im Folgenden werden die putativen Bindungsstellen jeder einzelnen Matrize mit denen der entsprechenden Zufallsmatrizen verglichen. Zunächst wurden die Anzahl der Zufallspositionen im jeweilig betrachteten Bereich in Prozente umgerechnet. Das heißt, es wurde bestimmt, wie groß der durchschnittliche Anteil der Zufallspositionen in jedem Bereich ist. Dann wurde entsprechend der Anteil der Transkriptionsfaktor-Bindungsstellen jeder Matrize in den gleichen Bereichen bestimmt. Anschließend wurden die Abweichungsfaktoren der matrizenbasierenden Bindungsstellen von den Zufallspositionen gebildet und grafisch aufgetragen. Diese erfolgte logarithmisch, da negative Abweichungsfaktoren durch diese Auftragungsart ebenso wie positive Abweichungsfaktoren sinnvoll dargestellt werden können. Im Folgenden werden die Ergebnisse für die pflanzlichen Matrizen dargestellt. In Abb. 43 bis Abb. 47 sind die Ergebnisse für den intergenischen Bereich, die erste 5'UTR, die ersten fünf Introns, die ersten vier Exons und die erste 3'UTR aufgeführt.

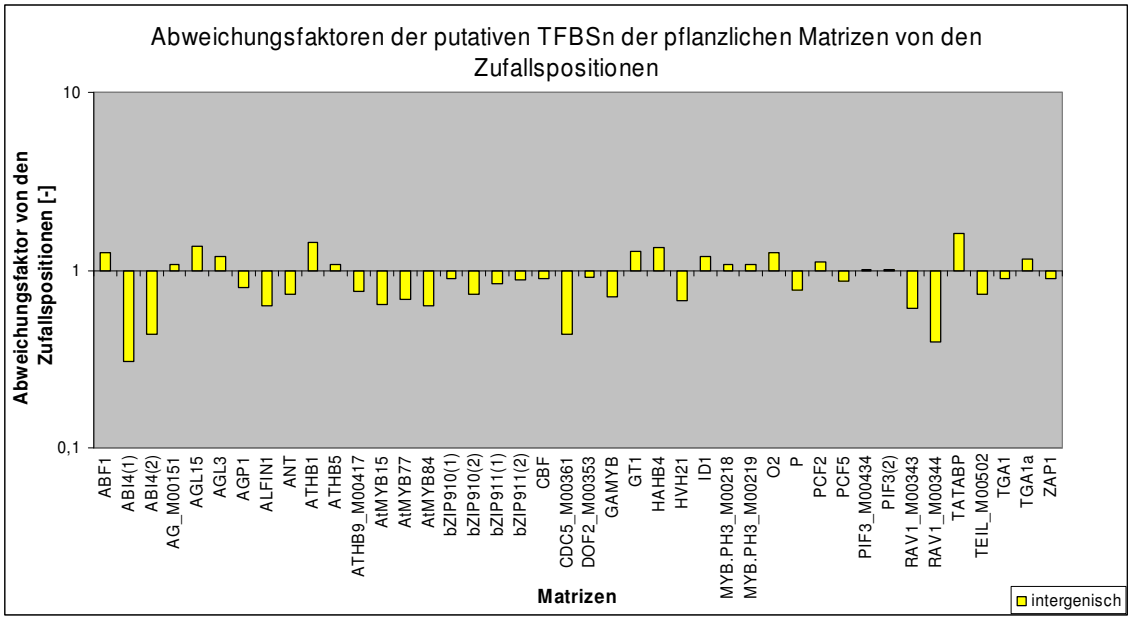


Abb. 43: Logarithmische Auftragung der Abweichungsfaktoren der putativen Transkriptionsfaktor-Bindungsstellen von den Zufallspositionen im intergenischen Bereich. Eine positive Abweichung wird bei 17 der 42 Matrizen (40,48 %) beobachtet. Die exakten Abweichungsfaktoren befinden sich im Anhang (Tabelle 43).

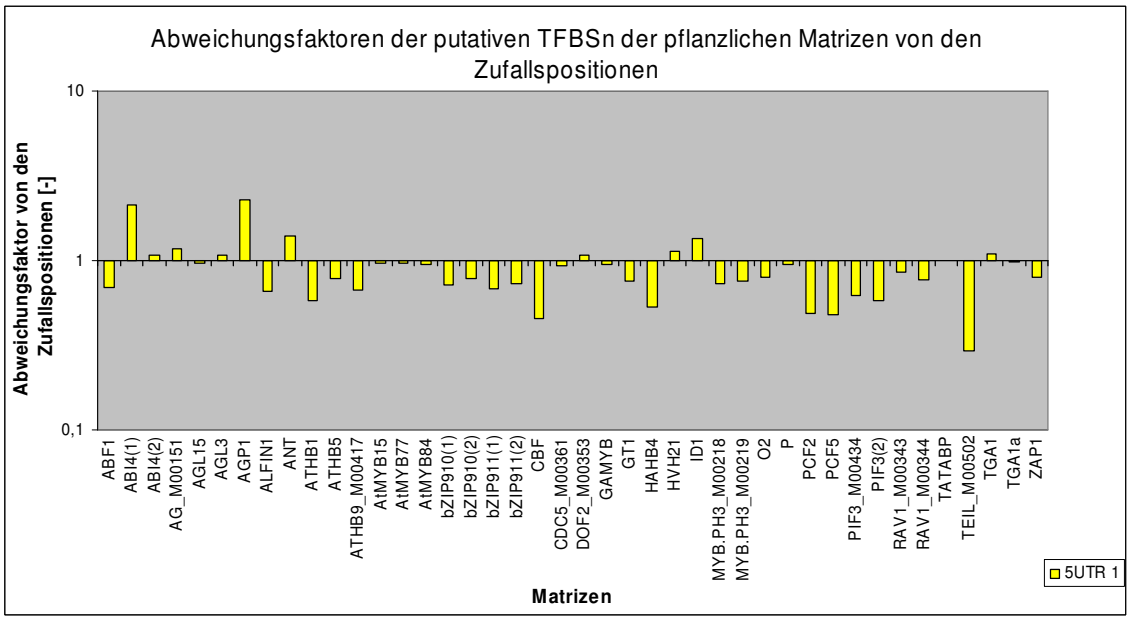


Abb. 44: Dargestellt sind die logarithmisch aufgetragenen Abweichungsfaktoren der putativen Transkriptionsfaktor-Bindungsstellen von den Zufallspositionen in der ersten 5'UTR. Eine positive Abweichung wird bei 11 der 42 Matrizen (26,19 %) beobachtet. Die exakten Werte befinden sich im Anhang (Tabelle 43).

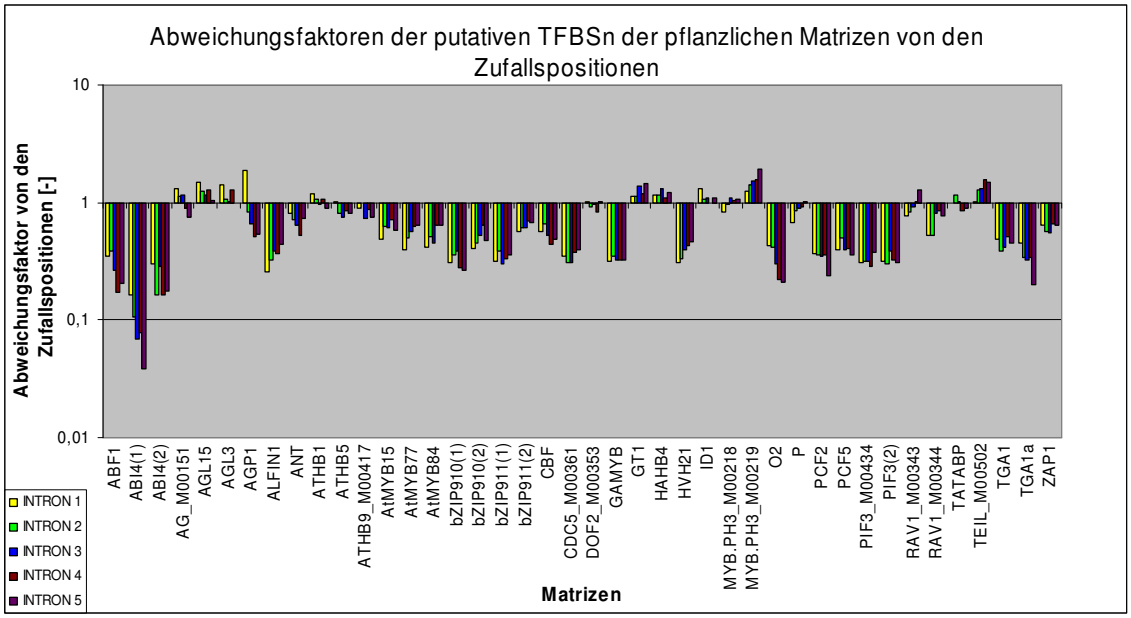


Abb. 45: Logarithmische Auftragung der Abweichungsfaktoren der putativen Transkriptionsfaktor-Bindungsstellen von den Zufallspositionen in den ersten fünf Introns. Auffällig ist, dass Matrizen, deren putative Bindungsstellen in einem bestimmten Intron eine Abweichungsrichtung aufweisen, diese vorwiegend auch in den weiteren betrachteten Introns beibehalten. Die exakten Werte befinden sich im Anhang (Tabelle 44).

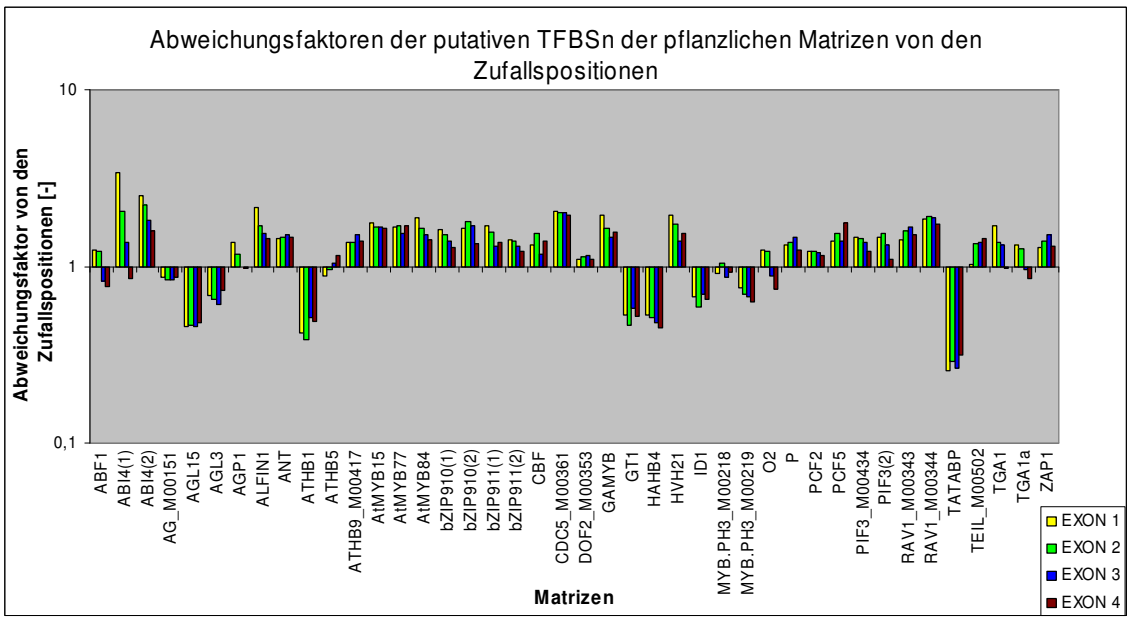


Abb. 46: Dargestellt sind die logarithmisch aufgetragenen Abweichungsfaktoren der putativen pflanzlichen Transkriptionsfaktor-Bindungsstellen von den Zufallspositionen in den ersten vier Exons. Innerhalb der betrachteten Exons bleibt die Abweichungsrichtung der putativen TFBSn einer Matrize in vielen Fällen gleich. Die exakten Werte sind im Anhang aufgeführt (Tabelle 43).

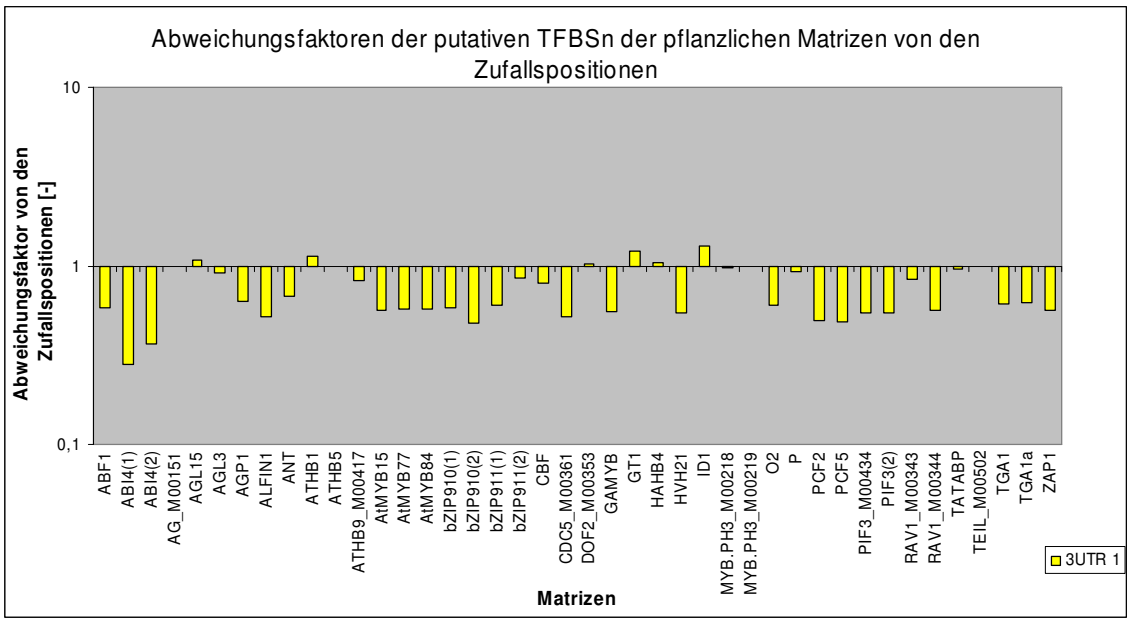


Abb. 47: Logarithmisch dargestellte Abweichungsfaktoren der putativen Transkriptionsfaktor-Bindungsstellen von den Zufallspositionen in der ersten 3'UTR. 6 der 42 Matrizen (14,29 %) weisen eine positive Abweichung im Vergleich zu den Zufallspositionen auf. Die exakten Werte befinden sich im Anhang (Tabelle 44).

Im intergenischen Bereich weisen 17 der 42 pflanzlichen Matrizen einen positiven Abweichungsfaktor von den Zufallspositionen auf (Abb. 43). Die positiven Abweichungsfaktoren sind im Schnitt kleiner als die negativen Abweichungsfaktoren. Die Maximalwerte betragen 1,6 für die positive und 0,6 für die negative Abweichung. Die Transkriptionsfaktoren, deren Daten positive Abweichungen aufweisen, gehören vorwiegend zu den Klassen MADS (3 von 3), bHLH (2 von 2), Trihelix (1 von 1), C2H2(Zn) (1 von 1), TBP (1 von 1), HD-ZIP (3 von 4), TCP (1 von 2), bZIP (3 von 8) sowie MYB (2 von 8).
Die Zufallsmatrizen der pflanzlichen Matrizen weisen nur bei 7 von 42 Matrizen positive Auffälligkeiten auf (Daten nicht gezeigt). Alle positiv auffälligen Zufallsmatrizen weisen ebenfalls positive Abweichungsfaktoren bei den entsprechenden pflanzlichen Matrizen auf.
10 pflanzliche Matrizen weisen unterschiedliche Abweichungsrichtungsfaktoren im Vergleich zu ihren Zufallsmatrizen auf. Diese sind ABF1, AGL15, AGL3, ATHB5, MYB.PH3_M00218, O2, PCF2, PIF3_M00434, PIF3(2) sowie TGA1a. Das deutet darauf hin, dass der AT-Gehalt der Matrizen nicht der alleinige Parameter für die Richtung des Abweichungsfaktors von den Zufallsmatrizen sein kann.
Nichtpflanzliche Matrizen weisen ebenfalls positive Abweichungsfaktoren von den Zufallspositionen auf, diese treten bei 4 von 27 Matrizen auf, wobei 3 Zufallsmatrizen der nichtpflanzlichen Matrizen diese ebenfalls zeigen (Daten nicht gezeigt). Diese 3

Zufallsmatrizen zeigen ebenfalls positive Abweichungsfaktoren bei ihren entsprechenden nichtpflanzlichen Matrizen, was darauf hindeutet, dass dieser Effekt aufgrund des AT-Gehalts der Matrizen resultieren könnte.

In der ersten 5'UTR weisen die dazugehörigen putativen Transkriptionsfaktor-Bindungsstellen der pflanzlichen Matrizen im Vergleich zu den Zufallspositionen in 11 Fällen positive Abweichungsfaktoren auf (Abb. 44). Die prinzipiellen Abweichungsfaktoren betragen Werte zwischen 2,3 und 0,29, wobei die Spanne größer als im intergenischen Bereich ist. 27 der 42 Zufallsmatrizen für die pflanzlichen Matrizen weisen ebenfalls positive Abweichungsfaktoren auf (Daten nicht gezeigt), was einen deutlichen Unterschied zum intergenischen Bereich darstellt. Dies deutet auf eine mögliche Gegenselektion von bestimmten DNA-Sequenzen, die durch die Matrizen beschrieben werden, hin. Bei den nichtpflanzlichen 27 Matrizen beträgt die Anzahl der positiven Abweichungen 13. Sogar 21 Zufallsmatrizen der nichtpflanzlichen Matrizen besitzen ebenfalls positive Abweichungsfaktoren (Daten nicht gezeigt).

In Introns zeigen die putativen Transkriptionsfaktor-Bindungsstellen der 42 pflanzlichen Matrizen im Vergleich zu den Zufallspositionen im ersten Intron 13, im zweiten und dritten je in 10, im vierten in 9 und im fünften in 10 Fällen positive Abweichungsfaktoren. Auffällig ist, dass die Richtung des Abweichungsfaktors oftmals in den fünf betrachteten Introns beibehalten wird. Insgesamt zeigen 11 Matrizen in Verlaufe der fünf Introns eine wechselnde Abweichungsrichtung, dies sind AG, AGL3, AGP1, ATHB1, ATHB5, DOF2, ID1, MYB.PH3_M00218, P RAV1_M00343 sowie TBP. Die Zufallsmatrizen der pflanzlichen Matrizen zeigen in 7 Fällen positive Abweichungsfaktoren in mindestens einem Intron, bei 4 Matrizen kommt es zu einer Richtungsänderung des Abweichungsfaktors im Verlauf der betrachteten Introns (Daten nicht gezeigt). Insgesamt 8 der 27 (29,63 %) nichtpflanzlichen Matrizen besitzen positive Abweichungsfaktoren in den untersuchten Introns, wobei es bei 6 Matrizen zu einer Änderung der Abweichungsrichtung kommt (Daten nicht gezeigt). Im Vergleich dazu weisen 3 Zufallsmatrizen der nichtpflanzlichen Matrizen positive Abweichungsfaktoren auf, 2 Matrizen zeigen veränderte Abweichungsrichtungen (Daten nicht gezeigt). Zusammengefasst kommt es bei Zufallsmatrizen seltener zu positiven Abweichungsfaktoren, wobei bei den nichtpflanzlichen Matrizen im Vergleich zu den pflanzlichen Matrizen etwa ähnlich viele Matrizen (in Prozenten) diese positiven Auffälligkeiten zeigen.

Die Analyse der ersten vier Exons ergab bei den putativen Transkriptionsfaktor-Bindungsstellen für die 42 pflanzlichen Matrizen im ersten Exon 31, im zweiten 32, im dritten

28 und im vierten Exon 26 positive Abweichungsfaktoren im Vergleich zu den Zufallspositionen. Die überwiegende Anzahl an Matrizen wiesen also positive Abweichungen auf. Bei 8 Matrizen kam es zu einer Änderung der Abweichungsrichtung von den Zufallspositionen. Im Vergleich dazu zeigten 34 Zufallsmatrizen für die pflanzlichen Matrizen positive Abweichungsfaktoren, wobei die positiven Abweichungen vorwiegend im ersten Exon auftreten. 16 der 42 Zufallsmatrizen wiesen im Verlauf der betrachteten Exons Abweichungsrichtungsänderungen auf (Daten nicht gezeigt). Bei den nichtpflanzlichen Matrizen wurden bei 25 Matrizen positive Abweichungsfaktoren festgestellt. 3 Matrizen ändern ihre Abweichungsrichtung von den Zufallspositionen (Daten nicht gezeigt). 24 der 27 Zufallsmatrizen der nichtpflanzlichen Matrizen wiesen in mindestens einem Exon positive Abweichungsfaktoren von den Zufallspositionen auf. Bei neun Matrizen kam es im Verlauf der untersuchten Exons zu einer Änderung der Abweichungsrichtung (Daten nicht gezeigt).

In der ersten 3'UTR wurden bei 6 pflanzlichen Matrizen positive Abweichungsfaktoren im Vergleich zu den Zufallspositionen festgestellt. Die Werte umfassten einen Bereich von 1,29 bei positiven Abweichungsfaktoren bis 0,28 bei negativen Faktoren, was größere Abweichungen im negativen Bereich beinhaltet. Im Vergleich dazu wiesen 7 Zufallsmatrizen der pflanzlichen Matrizen ebenfalls positive Abweichungsfaktoren auf, wobei die Faktoren etwa ähnliche Größenordnungen aufzeigten. Bei 3 pflanzlichen Matrizen wiesen die Zufallsmatrizen wie sie selbst ebenfalls positive Abweichungsfaktoren auf. Bei den nichtpflanzlichen Matrizen wurden bei 2 Matrizen positive Abweichungsfaktoren im Vergleich zu den Zufallspositionen festgestellt. Positive Abweichungsfaktoren bei den Zufallsmatrizen der nichtpflanzlichen Matrizen wurden in zwei Fällen ermittelt, wobei bei die Abweichungsrichtung bei einer Matrize mit der ihrer entsprechenden Zufallsmatrize übereinstimmte.

Um mögliche Aussagen über eine biologische Relevanz der Ergebnisse erhalten zu können, werden im Folgenden positive Abweichungsfaktoren der pflanzlichen Matrizen mit denen der anderen drei Matrizentypen (Zufallsmatrizen für die pflanzlichen Matrizen, nichtpflanzliche Matrizen sowie Zufallsmatrizen für die nichtpflanzlichen Matrizen) gegenübergestellt. Dadurch kann verglichen werden, ob der Anteil an positiven Abweichungen für die pflanzlichen Matrizen aufgrund der erwarteten detektierten Bindungsstellen im Vergleich zu den anderen drei Matrizentypen überwiegt. Eine Zusammenfassung der Ergebnisse ist in Abb. 48 dargestellt, in der die prozentuale Anzahl der pflanzlichen Matrizen, der Zufallsmatrizen der pflanzlichen Matrizen, der nichtpflanzlichen Matrizen sowie der Zufallsmatrizen der nichtpflanzlichen Matrizen mit einem positiven Abweichungsfaktor im

Vergleich zu den Zufallsmatrizen im intergenischen Bereich sowie in der ersten 5'UTR, den ersten fünf Introns, den ersten vier Exons und der 3'UTR visualisiert ist.

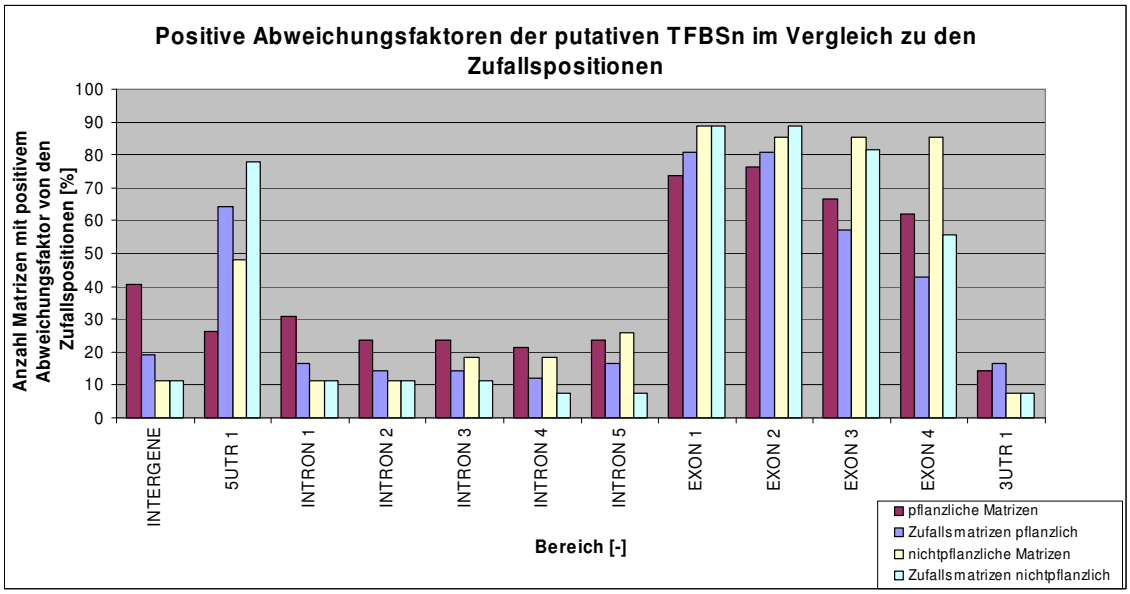


Abb. 48: Dargestellt ist der prozentuale Anteil der Matrizen mit positiven Abweichungsfaktoren von den Zufallspositionen in den untersuchten Bereichen innerhalb der Gene und im intergenischen Bereich. Die Ergebnisse für die pflanzlichen Matrizen, die Zufallsmatrizen für die pflanzlichen Matrizen, die nichtpflanzlichen Matrizen sowie deren Zufallsmatrizen sind zusammengefasst aufgeführt.

Im intergenischen Bereich ist eine deutliche Abweichung der pflanzlichen Matrizen im Vergleich zu den drei weiteren Matrizengruppen erkennbar. Es treten mehr pflanzliche Matrizen mit positiven Abweichungsfaktoren auf als dies bei den Vergleichsgruppen der Fall ist. Diese pflanzlichen Matrizen wurden protokolliert und sind in Tabelle 11 aufgeführt.

In der ersten 5'UTR kommen putative Transkriptionsfaktoren der pflanzlichen Matrizen weniger häufig vor als die der anderen drei Matrizentypen, was auf eine mögliche Gegenselektion hindeutet. Die pflanzlichen Matrizen mit positiven Abweichungsfaktoren wurden in Tabelle 11 aufgelistet.

Die putativen TFBSn der pflanzlichen Matrizen weisen in den ersten vier Introns mehr positive Abweichungsfaktoren auf als bei den Vergleichsgruppen, wobei ab dem dritten Intron die Anzahl der nichtpflanzlichen Matrizen mit positiven Abweichungsfaktoren tendenziell zunimmt. Die Abweichungsfaktoren der pflanzlichen Matrizen weisen im Vergleich zu den drei anderen Matrizentypen im fünften Intron kein einheitliches Ergebnis

auf. Der Anteil der Zufallsmatrizen für die nichtpflanzlichen Matrizen ist größer als der der pflanzlichen Matrizen. Insgesamt wurden die pflanzlichen Matrizen mit positiven Abweichungsfaktoren in den ersten vier Intons in die Tabelle 11 aufgenommen.

In den ersten beiden Exons kamen weniger pflanzliche Matrizen mit positiven Abweichungsfaktoren vor als dies bei den drei anderen Matrizentypen der Fall war. Generell war die Anzahl der Matrizen mit positiven Abweichungen jedoch hoch, sie lag bei über 70 % bei allen vier Matrizentypen. Aufgrund der weniger häufigen positiven Abweichungen bei pflanzlichen Matrizen könnte es sich um eine leichte Gegenselektion in Bezug auf bestimmte DNA-Sequenzen, die durch die Matrizen beschrieben werden, hindeuten. Protokolliert wurden daher in Tabelle 11 die Matrizen mit negativen Abweichungsfaktoren. Im Gegensatz zu den ersten beiden Exons wiesen die Exons 3 und 4 keine einheitliche Tendenz der pflanzlichen Matrizen im Vergleich zu den anderen drei Vergleichsgruppen auf. Im dritten Exon traten bei den Zufallsmatrizen für die pflanzlichen Matrizen weniger häufig positive Abweichungsfaktoren auf, die nichtpflanzlichen Matrizen und deren Zufallsmatrizen jedoch kamen häufiger vor als dies bei den pflanzlichen Matrizen der Fall war. Im vierten Exon traten positive Abweichungsfaktoren bei den pflanzlichen und nichtpflanzlichen Matrizen häufiger als bei ihren entsprechenden Zufallsmatrizen auf. Eine Protokollierung von auffälligen Matrizen in diesen beiden Exons wurde nicht durchgeführt.

Die Matrizen, deren putative Bindungsstellen in der ersten 3'UTR lokalisiert sind, zeigten kein einheitliches Ergebnis im Hinblick auf eine positive Abweichung von den Zufallspositionen. Die Zufallsmatrizen der pflanzlichen Matrizen wiesen mehr positive Abweichungsfaktoren auf als dies für die pflanzlichen Matrizen selbst zutraf. Die nichtpflanzlichen Matrizen sowie deren Zufallsmatrizen zeigten die gleiche Anzahl an positiven Abweichungsfaktoren, die jedoch geringer war als bei den erstgenannten Matrizentypen. Aufgrund des uneinheitlichen Ergebnisses wurde keine Protokollierung von pflanzlichen Matrizen in diesem Bereich durchgeführt.

Tabelle 11: Namen der pflanzlichen Matrizen, deren putativen Bindungsstellen aufgrund von Abb. 48: Dargestellt ist der prozentuale Anteil der Matrizen mit positiven Abweichungsfaktoren von den Zufallspositionen in den untersuchten Bereichen innerhalb der Gene und im intergenischen Bereich. Die Ergebnisse für die pflanzlichen Matrizen, die Zufallsmatrizen für die pflanzlichen Matrizen, die nichtpflanzlichen Matrizen sowie deren Zufallsmatrizen sind zusammengefasst aufgeführt.

Matrize	**Inter-genisch**	**5UTR 1**	**INTRON 1**	**INTRON 2**	**INTRON 3**	**INTRON 4**	**EXON 1**	**EXON 2**
ABF1	1,26	-	-	-	-	-	-	-

ABI4(1)	-	2,11	-	-	-	-	-	-
ABI4(2)	-	1,08	-	-	-	-	-	-
AG_M00151	1,07	1,18	1,3	1,12	1,14	-	0,86	0,83
AGL15	1,36	-	1,49	1,25	1,17	1,27	0,46	0,46
AGL3	1,18	1,08	1,4	1,08	1,02	1,27	0,68	0,65
AGP1	-	2,29	1,86	-	-	-	-	-
ANT	-	1,41	-	-	-	-	-	-
ATHB1	1,44	-	1,18	1,07	-	1,07	0,42	0,39
ATHB5	1,07	-	1,02	-	-	-	0,88	0,96
DOF2_M00353	-	1,07	1,01	-	-	-	-	-
GT1	1,27	-	1,12	1,12	1,37	1,17	0,53	0,47
HAHB4	1,35	-	1,17	1,14	1,32	1,09	0,53	0,51
HVH21	-	1,12	-	-	-	-	-	-
ID1	1,2	1,34	1,3	1,08	1,1	-	0,68	0,59
MYB.PH3_M00218	1,08	-	-	-	1,09	1,04	0,92	-
MYB.PH3_M00219	1,08	-	1,24	1,43	1,52	1,57	0,76	0,69
O2	1,26	-	-	-	-	-	-	-
PCF2	1,12	-	-	-	-	-	-	-
PIF3(2)	1	-	-	-	-	-	-	-
PIF3_M00434	1,01	-	-	-	-	-	-	-
RAV1_M00343	-	-	-	-	-	1,02	-	-
TATABP	1,6	1,01	1	1,15	1,01	-	0,26	0,29
TEIL_M00502	-	-	1,03	1,29	1,32	1,59	-	-
TGA1	-	1,1	-	-	-	-	-	-
TGA1a	1,16	-	-	-	-	-	-	-

Insgesamt wurden bei 26 der 42 Matrizen (61,9 %) Auffälligkeiten festgestellt. Aufgrund dieser Analysen gelten die putativen Transkriptionsfaktor-Bindungsstellen der pflanzlichen Matrizen in den in Tabelle 11 aufgeführten Bereichen als potenzielle Kandidaten für weiterführende Analysen mittels molekularbiologischer Methoden, um beispielsweise entsprechende DNA-Sequenzen, die in diesen Bereichen lokalisiert sind, überprüfen zu können.

3.4 Verteilung von TFBSn im intergenischen Bereich analysiert in 1 bp Schritten

Bisher wurden, wie in Kapitel 3.2 behandelt, Verteilungsanalysen der Zufallspositionen, der putativen TFBSn der pflanzlichen Matrizen, der Zufallsmatrizen der pflanzlichen Matrizen, der nichtpflanzlichen Matrizen sowie der Zufallsmatrizen der nichtpflanzlichen Matrizen im intergenischen Bereich durchgeführt. Diese Analysen wurden in 50 bp Schritten durchgeführt. Im weiteren Verlauf wurden nun verfeinerte Verteilungsanalysen im intergenischen Bereich in 1 bp Schritten unter Einsatz der TFBSitesTools (Kapitel 2.5.3) durchgeführt. Die durchschnittliche Länge des intergenischen Bereichs im *Arabidopsis thaliana* Genom beträgt 1793 bp (Kapitel 2.1.1). Da eine putative Transkriptionsfaktor-Bindungsstelle entweder upstream vom nächstgelegenen Gen oder aber downstream von diesem dazugehörigen Gen lokalisiert sein kann, ist die durchschnittliche Länge beidseitig zu betrachten. Ein negativer Wert ist das Kennzeichen für eine upstream-Lokalisierung der putativen Bindungsstelle (oder generell der Position) vom zugeordneten Gen. Es sei noch darauf hingewiesen, dass Bindungsstellen, die im downstream-Bereich auftreten, räumlich von denen im upstream-Bereich durch die Länge des Gens getrennt sind. Downstream bedeutet also in diesem Fall downstream vom Ende des Gens aus gesehen.

Zunächst wurden die 42 x 7500 Zufallspositionen im intergenischen Bereich grafisch aufgetragen (Daten nicht gezeigt). Die Schrittweite betrug 1 bp. Die betrachtete Distanz zum nächstgelegenen Gen wurde von -1793 bp bis -1 bp upstream und von 1 bp bis 1793 bp downstream untersucht. Hierbei zeigten alle Grafiken der Zufallspositionen eine Anreicherung von Positionen im upstream-Bereich, die mit abnehmender Distanz zunahm. Im downstream-Bereich waren hingegen deutlich weniger Zufallspositionen lokalisiert (Daten nicht gezeigt).

Nach der grafischen Auswertung wurde der Anteil der 315000 Zufallspositionen an den einzelnen Distanzen relativ zu den Genen addiert und in Prozente umgerechnet, wonach relativ zur Gesamtzahl an Zufallspositionen ein Wert gebildet wurde. Ausgehend von den Zufallspositionen erfolgte für jede einzelne Distanz die Berechnung des Abweichungsfaktors der putative(n) Bindungsstelle(n) jeder pflanzlichen Matrize, Zufallsmatrize der pflanzlichen Matrize, nichtpflanzlichen Matrize sowie Zufallsmatrize der nichtpflanzlichen Matrize von denen der Zufallspositionen. Anschließend wurden Überlagerungsgrafiken mit den Abweichungsfaktoren einer pflanzlichen bzw. nichtpflanzlichen Matrize und ihrer jeweiligen Zufallsmatrize angefertigt. Die Werte wurden logarithmisch aufgetragen. Beispielhaft ist eine dieser Grafiken in Abb. 49 dargestellt. Auffällig ist, dass es besonders im upstream-Bereich zu negativen Abweichungsfaktoren im Vergleich zu den Zufallspositionen kam, und zwar galt dies sowohl für die putativen Bindungsstellen der pflanzlichen Matrize CBF als auch für die

putativen Bindungsstellen ihrer dazugehörigen Matrize, randomCBF. Die Beobachtung der negativen Abweichungsfaktoren besonders im upstream-Bereich konnte bei allen Überlagerungsgrafiken, also auch bei denen für die nichtpflanzlichen Matrizen, festgestellt werden. Generell wurde diese negative Abweichung mit abnehmender Distanz zum Gen tendenziell größer.

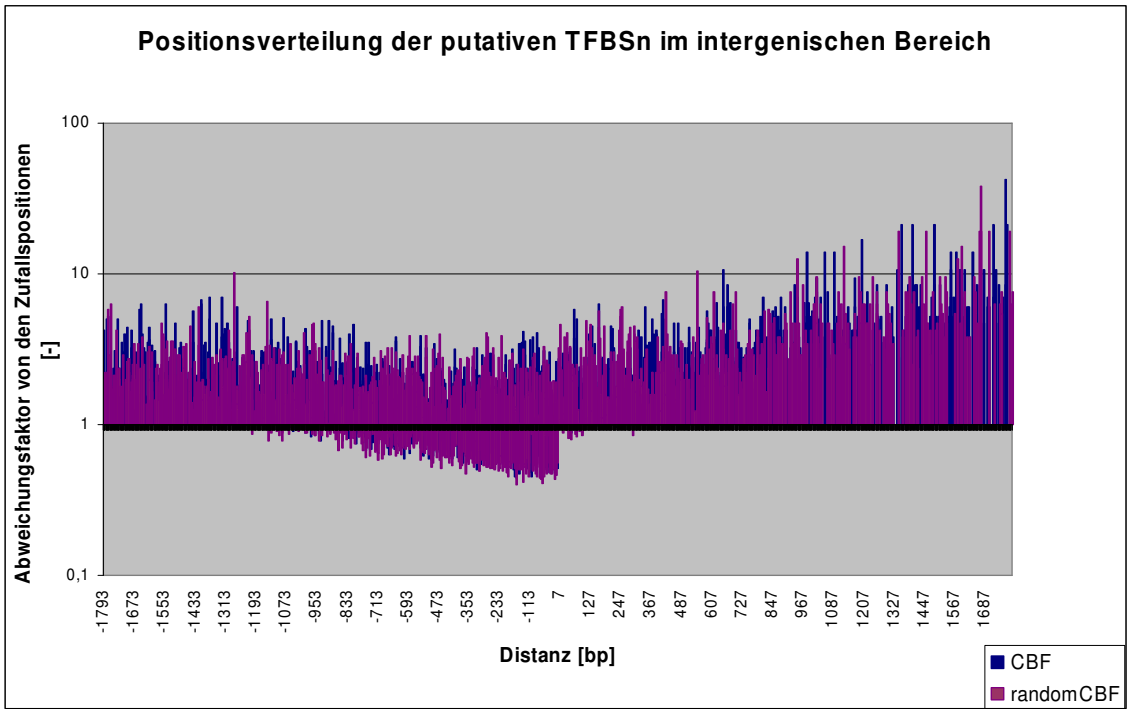


Abb. 49: Überlagerungsgrafik der putativen Transkriptionsfaktor-Bindungsstellen von CBF und der entsprechenden Zufallsmatrize im intergenischen Bereich, analysiert in 1 bp Schritten. Aufgetragen sind die Abweichungsfaktoren von den Zufallspositionen an der jeweiligen Position. Im upstream-Bereich von Genen kam es gehäuft zu negativen Abweichungsfaktoren. Diese negative Abweichungen traten sowohl bei putativen TFBSn der pflanzlichen Matrize als auch bei denen ihrer entsprechenden Zufallsmatrize auf.

Dies kann auf einen prinzipiellen biologischen Einfluss der Nukleotidkomposition in dem Bereich hindeuten, was nicht allein vom AT-Gehalt abhängig ist, da dies bei allen vier matrizenbasierenden Versuchsgruppen auftritt. Daher wurden die Analysen unter weiter verfeinerten Bedingungen wiederholt und in zwei Gruppen eingeteilt. Putative Bindungsstellen wurden mittels der TFBSitesTools ihrem jeweiligen Gen zugeordnet, das der Bindungsstelle am nächsten liegt. Hierbei lässt sich eine Unterscheidung zwischen Genen, die einen annotierten Transkriptionsstart haben und Genen mit nur einem annotierten Translationsstart besitzen, durchführen. Im Folgenden wurden die putativen Bindungsstellen,

deren dazugehöriges Gen einen bekannten Transkriptionsstart aufweist sowie für deren zugeordnetes Gen kein Transkriptionsstart annotiert ist, getrennt voneinander betrachtet.

Zunächst wurden Verteilungsanalysen der Zufallspositionen im intergenischen Bereich für beide Fälle durchgeführt und mit denen für das vereinfachte Kriterium des intergenischen Bereichs, bei dem putative TFBSn mit einbezogen wurden, deren dazugehöriges Gen einen Transkriptionsstart aufwies oder eben nicht, verglichen. Hierbei war bei allen 84 Grafiken (je 42 für die Analysen mit einem vorhandenen Transkriptionsstart und mit nur einem annotierten Translationsstart) ebenfalls eine Anreicherung der Positionen im upstream-Bereich zu verzeichnen, die Grafiken wiesen ähnliche Verteilungsmuster im Vergleich zu dem vereinfachten Kriterium, auf (Daten nicht gezeigt). Anschließend wurden, analog zu den oben genannten Analysen, für alle betrachteten Distanzen im Bereich von -1793 bp bis 1793 bp bei den matrizenbasierenden Datensätzen Abweichungsfaktoren von den Zufallspositionen gebildet und die Werte analog zu Abb. 49 grafisch aufgetragen. In allen Fällen waren wieder negative Abweichungsfaktoren bei den pflanzlichen Matrizen, den Zufallsmatrizen der pflanzlichen Matrizen, der nichtpflanzlichen Matrizen sowie deren Zufallsmatrizen im upstream-Bereich zu verzeichnen. Die Bereiche, in denen die negative Abweichung auftraten, wurden für alle vier Versuchsgruppen quantifiziert und sind in Tabelle 12, unter Berücksichtigung, dass die dazugehörigen Gene der putativen Bindungsstellen einen Transkriptionsstartpunkt aufweisen, und Tabelle 13, für den Fall, dass die dazugehörigen Gene der putativen Bindungsstellen lediglich einen annotierten Translationsstart besitzen, aufgeführt. Die Mittelwerte sowie die Fehler der Mittelwerte wurden wie in Kapitel 2.6.1 aufgeführt, berechnet.
Wie aus Tabelle 12 ersichtlich, lagen die Mittelwerte der maximalen Distanz zum nächstgelegenen Gen mit negativen Abweichungsfaktoren für die putativen TFBSn der pflanzlichen Matrizen, der Zufallsmatrizen der pflanzlichen Matrizen sowie bei denen der nichtpflanzlichen Matrizen nah beieinander. Lediglich bei den putativen TFBSn der Zufallsmatrizen der nichtpflanzlichen Matrizen lag die maximale Distanz um etwa 50 Nukleotide näher am nächstgelegenen Gen. Die Mittelwerte der minimalen Distanz zum nächstgelegenen Gen mit negativen Abweichungsfaktoren von den Zufallspositionen der vier Versuchsgruppen lagen alle zwischen -3,81 bp bis -2,2 bp und sind daher recht einheitlich. Insgesamt konnte ein gesamter Mittelwert der maximalen und minimalen Distanz zum nächstgelegenen Gen von -405 bp bis -3 bp verzeichnet werden.
Das Auftreten von negativen Abweichungsfaktoren von den Zufallspositionen trat bei allen matrizenbasierenden Versuchsgruppen auf, unabhängig davon, ob reale Bindungsstellen durch eine Suche nach TFBSn erwartet wurden oder nicht.

Tabelle 12: Mittelwerte der maximalen und minimalen Distanzen mit negativen Abweichungsfaktoren von den Zufallspositionen relativ zum nächstgelegenen Transkriptionsstart. Es wurden nur putative TFBSn in die Analyse miteinbezogen, deren dazugehörige Gene einen annotierten Transkriptionsstart aufwiesen. Das Vorzeichen des Mittelwerts gibt die Lokalisierung der Distanz relativ zum Gen an, ein negativer Wert bedeutet, dass die dazugehörigen putativen TFBSn upstream von den Genen liegen. Die Distanzwerte sind in bp angegeben.

	Mittelwert der maximalen Distanz mit neg. Abweichungsfaktoren	**Fehler des Mittelwerts (max. Distanz)**	**Mittelwert der minimalen Distanz mit neg. Abweichungsfaktoren**	**Fehler des Mittelwerts (min. Distanz)**
pflanzliche Matrizen	-417,48	7,63	-2,6	0,45
Zufallsmatrizen (pflanzlich)	-422,19	16,13	-2,6	0,72
nichtpflanzliche Matrizen	-413,11	14,59	-3,81	0,74
Zufallsmatrizen (nichtpflanzlich)	-363,59	12,95	-2,22	0,54

Tabelle 13: Mittelwerte der Distanzen mit negativen Abweichungsfaktoren von den Zufallspositionen relativ zum nächstgelegenen Translationsstart. Es wurden nur putative TFBSn in die Analyse miteinbezogen, deren dazugehörige Gene keinen annotierten Transkriptionsstart, sondern nur einen Translationsstartpunkt, aufwiesen. Das Vorzeichen des Mittelwerts gibt die Lokalisierung der Distanz relativ zum Gen an, ein negativer Wert bedeutet, dass die dazugehörigen putativen TFBSn upstream von den Genen liegen. Die Distanzwerte sind in bp angegeben.

	Mittelwert der maximalen Distanz mit neg. Abweichungsfaktoren	**Fehler des Mittelwerts (max. Distanz)**	**Mittelwert der minimalen Distanz mit neg. Abweichungsfaktoren**	**Fehler des Mittelwerts (min. Distanz)**
pflanzliche Matrizen	-367,05	14,08	-12,81	2,11
Zufallsmatrizen (pflanzlich)	-377,95	19,76	-11,64	2,44
nichtpflanzliche Matrizen	-338,63	14,5	-11,44	1,77
Zufallsmatrizen (nichtpflanzlich)	-347,22	8,78	-12,74	2,04

Wurden putative Transkriptionsfaktor-Bindungsstellen in die Analysen miteinbezogen, deren dazugehöriges Gen keinen Transkriptionsstart, aber einen Translationsstart aufweisen (Tabelle 13), traten ebenfalls negative Abweichungsfaktoren von den Zufallspositionen auf. Der Mittelwert der maximalen Distanz mit negativen Abweichungsfaktoren wies Werte zwischen -377,95 bp bis -338,63 bp, das heißt, die maximale Distanz war etwas näher am nächstgelegenen Gen gelegen als bei den putativen TFBSn, deren dazugehöriges Gen einen Transkriptionsstartpunkt aufweist. Die Mittelwerte der minimalen Distanz mit negativen Abweichungsfaktoren von den Zufallspositionen lagen nah beieinander. Insgesamt konnten Gesamtmittelwerte aller vier Versuchsgruppen von -358 bp bis -12 bp festgestellt werden.

Da alle matrizenbasierenden Versuchsgruppen negative Abweichungsfaktoren von den Zufallspositionen zeigten, deutet dies auf eine nicht zufällige Nukleotidkompositionen hin. Möglicherweise könnte der Bereich von ca. -405 bp bis -3 bp auf eine bevorzugte Lokalisierung von Transkriptionsfaktor-Bindungsstellen im Allgemeinen hindeuten.

Im Folgenden wurden die Überlagerungsgrafiken, siehe beispielhaft Abb. 49, auf ein unterschiedliches Verteilungsmuster der putativen TFBSn der realen Matrize im Vergleich zu denen der entsprechenden Zufallsmatrize betrachtet. Hierbei wurde eine Beschränkung auf den oben definierten Bereich, -358 bp bis -12 bp bzw. -405 bp bis -3 bp, in dem es möglicherweise zu einer Anhäufung von regulatorischen DNA-Sequenzen kommen könnte, durchgeführt. Es wurden sowohl die Überlagerungsgrafiken für die pflanzlichen Matrizen wie auch die für die nichtpflanzlichen Matrizen auf Auffälligkeiten hin untersucht. Bei dieser Analyse wurden, wie oben aufgeführt, das Vorhandensein eines annotierten Transkriptionsstartpunkts des entsprechenden Gens der putativen TFBS und das Fehlen dieses Punktes gesondert betrachtet. Im Folgenden wird zunächst das Vorhandensein eines annotierten Transkriptionsstarts des zu der putativen TFBS gehörenden Gens gefordert.

In den Überlagerungsgrafiken der pflanzlichen Matrizen und ihrer entsprechenden Zufallsmatrizen wurden drei Beobachtungen gemacht. Es kam in einigen Fällen zu nicht auffälligen Überlagerungen, das heißt, dass sich die Anzahl der putativen Transkriptionsfaktor-Bindungsstellen der beiden Matrizen in den betrachteten Distanzen auf ähnlichen Niveaus bewegt, Abb. 49. Bei einigen Matrizen konnten jedoch Auffälligkeitsbereiche festgestellt werden. Innerhalb dieser, die im upstream-Bereich zu den entsprechenden Genen lagen, zeigten die putativen TFBSn der pflanzlichen Matrize eine Anreicherung bei Distanzbereichen. Diese Anreicherung fehlte bei den entsprechenden Zufallsmatrizen. Abb. 50 verdeutlicht diese Beobachtung anhand der putativen TFBSn des

Transkriptionsfaktors bZIP910. Insgesamt waren diese Auffälligkeitsbereiche bei 11 der 42 pflanzlichen Matrizen vorhanden. Faktoren einer Transkriptionsfaktor-Klasse erkennen oft ähnliche DNA-Sequenzen, weil die DNA-Bindungsdomänen der Faktoren sehr ähnlich sind. Bei den matrizenbasierenden Datensätzen bezüglich der Klassen bZIP, bHLH und TCP konnte eine Korrelation zwischen der Auffälligkeit und der Klasse festgestellt werden. Die meisten der eingesetzten Matrizen dieser drei Transkriptionsfaktor-Klassen zeigten deutliche Anreicherungen von putativen TFBSn im upstream-Bereich der zugeordneten Genen (Tabelle 14). Das deutet darauf hin, dass Transkriptionsfaktoren dieser drei Klassen häufig putative Bindungsstellen im upstream-Bereich von Genen aufweisen.

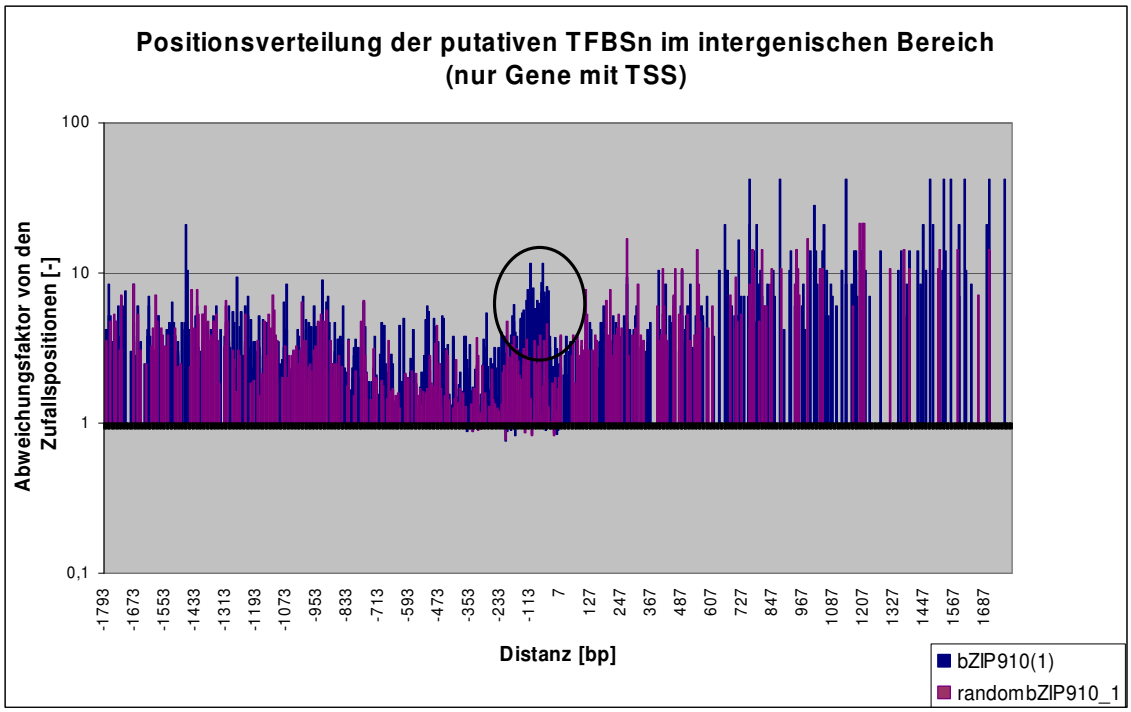


Abb. 50: Dargestellt ist die Positionsverteilung der putativen TFBSn von bZIP910(1) sowie die der entsprechenden Zufallsmatrize relativ zum nächstgelegenen Gen. Hierbei wurden nur jene putativen TFBSn für die Analyse berücksichtigt, deren dazugehörige Gene einen annotierten Transkriptionsstartpunkt aufweisen. Der schwarze Kreis verdeutlicht den Auffälligkeitsbereich, in dem die Anzahl der putativen pflanzlichen TFBSn deutlich größer als bei der entsprechenden Zufallsmatrize ist. Des Weiteren sind in diesem Bereich auch einzelne Distanzen zu erkennen, bei denen noch mehr putative TFBSn der realen Matrize auftreten als im übrigen Bereich, innerhalb dessen die Auffälligkeit festgestellt wurde.

Eine weitere Auffälligkeit war das Auftreten von einzelnen Distanzen im upstream-Bereich von Genen, bei denen deutlich mehr putative TFBSn der pflanzlichen Matrizen vorkamen als bei den entsprechenden Zufallsmatrizen. Beispielhaft sei wieder auf Abb. 50 verwiesen, in

der diese Anreicherung im upstream-Bereich mit einem schwarzen Kreis gekennzeichnet ist. Eine Anreicherung dieser Art wurde genau dann als solche angesehen, wenn im gesamten untersuchten Distanzbereich bei einer bestimmten Distanz mehr putative pflanzliche TFBSn auftraten als bei den putativen TFBSn der entsprechenden Zufallsmatrize bei irgendeiner Distanz in dem untersuchten Bereich, was bei 11 Matrizen der Fall war (Tabelle 15).

Tabelle 14: Anteil pflanzlicher Matrizen, die Auffälligkeitsbereiche aufweisen. Das Augenmerk ist auf die Transkriptionsfaktor-Klassen gerichtet. In die Analyse wurden nur putative TFBSn mit einbezogen, deren dazugehörige Gene einen annotierten Transkriptionsstart aufwiesen. In bestimmten Bereichen ist die Anzahl putativer Transkriptionsfaktor-Bindungsstellen der realen Matrize im Vergleich zu ihrer entsprechenden Zufallsmatrize größer.

TF-Klasse	**Anteil auffälliger Matrizen [%]**	**Anteil auffälliger Matrizen/ Gesamtanzahl Matrizen der Klasse**
bHLH	100	2/2
TCP	100	2/2
TBP	100	1/1
bZIP	75	6/8

Tabelle 15: Anzahl pflanzlicher Matrizen, die bei bestimmten Distanzen eine größere Anzahl putativer TFBSn aufwiesen als dies bei den entsprechenden Zufallsmatrizen im gesamten untersuchten Bereich der Fall war. In die Analyse wurden nur putative TFBSn mit einbezogen, deren dazugehörige Gene einen annotierten Transkriptionsstart aufwiesen.

Matrize	**TF-Klasse**
bZIP910(2)	bZIP
bZIP911(1)	bZIP
bZIP911(2)	bZIP
TGA1	bZIP
TGA1a	bZIP
CBF	CBF
HVH21	HD-Knotted
ATHB5	HD-ZIP
AtMYB15	MYB
AtMYB77	MYB
GAMYB	MYB

Bei den nichtpflanzlichen Matrizen konnten keine Auffälligkeitsbereiche festgestellt werden. Es konnten aber Matrizen identifiziert werden, die an bestimmten Distanzen mehr putative TFBSn aufwiesen als ihre entsprechenden Zufallsmatrizen im gesamten analysierten Bereich. Jedoch war die Anzahl dieser Matrizen im Verhältnis geringer (Tabelle 16) als dies bei den pflanzlichen Matrizen der Fall war. Dies deutet darauf hin, dass besonders das Auftreten von Auffälligkeitsbereichen bei den pflanzlichen Matrizen auf eine biologische Relevanz hindeuten könnte.

Tabelle 16: Anzahl nichtpflanzlicher Matrizen, die bei bestimmten Distanzen eine größere Anzahl putativer TFBSn aufwiesen als dies bei den entsprechenden Zufallsmatrizen im gesamten untersuchten Bereich der Fall war. In die Analyse wurden nur putative TFBSn mit einbezogen, deren dazugehörige Gene einen annotierten Transkriptionsstart aufwiesen.

Matrize	**TF-Klasse**
AREB6_M00415	CH+homeo
Brachyury_M00150	T- Box
COUP_M00158	CC (rec)
FOXA4a_M00267	Fork head
GCM_M00270	GCM
HAP3_M00288	histone fold
LAC9_M00207	C6
Pax-5_M00143	paired

Wurden nur putative Transkriptionsfaktor-Bindungsstellen der Matrizen in die Analysen miteinbezogen, deren dazugehörige Gene keinen Transkriptionsstartpunkt aufwiesen, sondern bei denen nur der Translationsstartpunkt bekannt war, war nur bei der Matrize des Faktors TBP ein Auffälligkeitsbereich festzustellen. Dieser lag bei -125 bis -53 bp im upstream-Bereich der dazugehörigen Gene (Daten nicht gezeigt). 17 Matrizen wiesen einzelne Distanzen auf, bei denen die Anzahl der putativen TFBSn größer war als dies bei den entsprechenden Zufallsmatrizen in dem gesamten analysierten Bereich der Fall war (Tabelle 17).
Auffällig hierbei waren die matrizenbasierenden Datensätze der Transkriptionsfaktor-Klassen bZIP und HD-ZIP. Es wurden mehrere Matrizen von Faktoren dieser Klassen für die Analysen eingesetzt und die meisten zeigten diese Auffälligkeit.

Bei den nichtpflanzlichen Matrizen kamen keine Distanzbereiche vor, innerhalb derer die Anzahl der putativen TFBSn größer war als bei den entsprechenden Zufallsmatrizen. Auch die Anzahl der Matrizen, die an bestimmten Distanzen mehr putative TFBSn aufwiesen als

im gesamten betrachteten Bereich der entsprechenden Zufallsmatrizen (Tabelle 18), war deutlich geringer als bei den pflanzlichen Matrizen.

Tabelle 17: Anteil pflanzlicher Matrizen, die an bestimmten Distanzen mehr putative Bindungsstellen aufwiesen als die entsprechenden Zufallsmatrizen im gesamten untersuchten Distanzbereich. Das Augenmerk ist auf die Transkriptionsfaktor-Klassen gerichtet. In die Analyse wurden nur putative TFBSn einbezogen, deren dazugehörige Gene keinen annotierten Transkriptionsstart aufwiesen, sondern nur einen Translationsstart besaßen.

TF-Klasse	Anteil auffälliger Matrizen [%]	Anteil auffälliger Matrizen/ Gesamtanzahl Matrizen der Klasse
TBP	100	1/1
CBF	100	1/1
WRKY	100	1/1
bZIP	87,5	7/8
HD-ZIP	75	3/4
TCP	50	1/2
MYB	37,5	3/8

Tabelle 18: Anzahl nichtpflanzlicher Matrizen, die bei bestimmten Distanzen eine größere Anzahl putativer TFBSn aufwiesen als dies bei den entsprechenden Zufallsmatrizen im gesamten untersuchten Bereich der Fall war. In die Analyse wurden nur putative TFBSn mit einbezogen, deren dazugehörige Gene lediglich einen annotierten Transkriptionsstart aufwiesen.

Matrize	TF-Klasse
abaA_M00027	TEA
HAP3_M00288	histone fold
Pax-2_M00098	paired
Pax-5_M00143	paired
Pax-5_M00144	paired
Pax-6_M00097	paired-homeo

Zusammengefasst lässt sich feststellen, dass generell weniger Auffälligkeiten bei den nichtpflanzlichen Matrizen zu verzeichnen waren als bei den pflanzlichen Matrizen. Wurde bei den Analysen berücksichtigt, dass die dazugehörigen Gene der putativen TFBSn einen annotierten Transkriptionsstartpunkt aufwiesen, waren bei den pflanzlichen Matrizen Distanzbereiche festzustellen, innerhalb derer deutlich mehr putative TFBSn auftraten als bei den entsprechenden Zufallsmatrizen in diesen Bereichen. Diese Beobachtung konnte bei

den nichtpflanzlichen Matrizen nicht festgestellt werden, die keine Auffälligkeitsbereiche aufwiesen.

Das Auftreten von einzelnen Distanzen, an denen die Anzahl der putativen pflanzlichen bzw. nichtpflanzlichen TFBSn auffällig war, wurde näher untersucht. Die entsprechenden Gene der putativen TFBSn an den auffälligen einzelnen Distanzen wurden extrahiert und deren Funktion wurde untersucht. Hierbei konnte festgestellt werden, dass die Anzahl der Gene, die für Transposons, Pseudogene oder hypothetische Proteine kodieren, bei einem geforderten bekannten Transkriptionsstartpunkt des zur putativen TFBS gehörenden Gens deutlich verringert werden konnte (Daten nicht gezeigt).

Bislang wurden die Zufallsmatrizen als Hintergrundwerte verwendet und das Auftreten von putativen TFBSn von pflanzlichen bzw. nichtpflanzlichen Matrizen mit denen ihrer entsprechenden Zufallsmatrizen verglichen. Im weiteren Verlauf der Analysen wurde nur noch auf die pflanzlichen bzw. nichtpflanzlichen Matrizen und deren Daten eingegangen. Als weiteren Ansatzpunkt für das Auffinden von Auffälligkeiten wurde ein statistisches Kriterium eingesetzt. Auf die Ergebnisse wird im nächsten Kapitel eingegangen.

3.5 Einsatz des P-Wertes zur Bestimmung von signifikanten Anhäufungen von putativen TFBSn

Ausgehend von den Positionsanalysen relativ zu den einzelnen Bereichen innerhalb von Genen sowie im intergenischen Bereich, analysiert in 1 bp Schritten, wurden in Kapitel 3.4 Überlagerungsgrafiken erstellt, um Gemeinsamkeiten und Unterschiede zwischen den Zufallspositionen und den realen Matrizen im Vergleich zu ihren entsprechenden Zufallsmatrizen aufzuzeigen.

Ein anderer Ansatzpunkt zur Identifizierung von Auffälligkeiten ist die Verwendung eines statistischen Kriteriums, welches im Folgenden eingesetzt wurde. Für jede einzelne Distanz in den analysierten Bereichen wurde für die putativen TFBSn jeder einzelnen Matrize bzw. für die Zufallspositionen ein P-Wert berechnet (Kapitel 2.6.3). Das Signifikanzniveau wurde auf P = 0,05 festgelegt, das heißt, eine Signifikanz lag genau dann vor, wenn der berechnete P-Wert kleiner das Signifikanzniveau war.

Im Folgenden werden weitergehende Analysen mit den 10 signifikantesten P-Werten jeder Distanz, beschränkt auf die pflanzlichen Matrizen, beschrieben.

3.5.1 Beispielhafte Betrachtung der je 10 signifikantesten Anhäufungen in den einzelnen Bereichen innerhalb von Genen und im intergenischen Bereich

Für jede einzelne Distanz, an der putative Bindungsstellen vorkamen, wurde zunächst der P-Wert berechnet. Je niedriger der P-Wert ist, umso größer ist die Wahrscheinlichkeit, dass das Auftreten von putativen Bindungsstellen an der Position im Genom nicht zufällig ist. Prinzipiell gilt, dass das Auftreten von signifikanten P-Werten auf eine mögliche Koregulation der den putativen Transkriptionsfaktor-Bindungsstellen zugeordneten Gene hindeutet. Im Folgenden wurden die jeweils 10 signifikantesten, also niedrigsten, P-Werte für jeden untersuchten Bereich identifiziert. Berücksichtigt wurden hierbei nur die P-Werte, die zu den pflanzlichen Matrizen gehören. Die analysierten Bereiche sind die erste 5'UTR, die ersten fünf Introns, die ersten vier Exons, die erste 3'UTR sowie der intergenische Bereich, unterteilt in TFBSn für Gene mit einem annotierten Transkriptionsstartpunkt sowie für Gene mit lediglich einem bekannten Translationsstartpunkt.

Unter Einsatz der TFBSitesTools (Kapitel 2.5.3) wurden die Gennamen, die zu der Signifikanz gehören, extrahiert. Anschließend erfolgte die Extraktion der Genfunktionen mit der Software BuildGeneQuery (Kapitel 2.5.4) aus der AthaMap Datenbank. Die Gennamen wurden danach bei Genevestigator (Kapitel 2.9) eingesetzt, um Expressionsprofile dieser Gene zu erhalten. Erhalten wurden bei Genevestigator drei Grafiken, die Informationen der Genexpression aufgeteilt in Anatomie (A, Anatomy), Entwicklung (D, Development) sowie

Stimulus (S, Stimulus) enthielten. Grafiken bezüglich der Anatomie und der Entwicklung beinhalten die prinzipielle Expression der Gene bzw. die nicht vorhandene Expression. Der Stimulus gibt eine Induzierung bzw. Reprimierung der Genexpression durch Zuführung verschiedener Chemikalien, biotischem oder abiotischem Stress, Hormonen usw. an.

Die 10 signifikantesten P-Werte in der ersten 5'UTR sind in Tabelle 19 aufgeführt. Insgesamt konnten putative TFBSn von 6 verschiedenen Faktoren festgestellt werden. ABI4 war sogar viermal vertreten. Überraschend ist, dass der Transkriptionsfaktor TBP ebenfalls auffällig war. Die von dem Faktor erkannte TATA-box befindet sich im intergenischen Bereich. In allen Fällen konnte eine Koexpression der eingegebenen Gene, belegt durch Genevestigator, festgestellt werden. Als Gemeinsamkeit wurde angesehen, wenn z.B. in mindestens einem der untersuchten Gewebe eine Expression aller Gene beobachtet wurde bzw. ausblieb. Am Beispiel von AtMYB77 sind die Genevestigor-Grafiken in Abb. 51 dargestellt. Die putativen Zielgene sind At3g22960, At3g23090, At3g25860, At4g26210, At5g55510, wobei At5g55510 nicht in Genevestigator vorhanden war. In der Entwicklung (D) trat eine gemeinsame Expression in reifen Schoten (mature siliques) und im Keimling (seedling) auf. Bei der Anatomie (A) war eine Expression der Gene im Keimling (seedling), sowie einiger Komponenten des Keimlings wie das Hypocotyl (hypocotyl) und die Keimwurzel (radicle) zu verzeichnen. Des Weiteren konnte im Blütenstand (inflorescence), den Schoten (silique), im Stängel (stem) sowie in der Wurzel (roots) und einiger Bereiche innerhalb der Wurzel die gemeinsame Expression der putativen Zielgene verzeichnet werden. Eine fehlende Expression der Gene konnte im Pollen (pollen) aufgezeigt werden. Einen Einfluss auf Stimuli konnte bei Cycloheximid (cycloheximide), Seneszenz (senescence) und osmotischem Stress (osmotic stress) in Form einer Herunterregulierung der Gene und im Falle von *Agrobacterium tumefaciens* einer Hochregulierung verzeichnet werden. Eine unterschiedliche Regulation trat unter Einfluss von Syringolin (syringolin) auf. Wurde die Expression des Transkriptionsfaktors mit der Expression der putativen Zielgene verglichen, konnten übereinstimmende Regulationen in der ausgereiften Schote, im Hypocotyl, im Keim und im Pollen festgestellt werden. Diese übereinstimmenden Ergebnisse deuten stark auf eine mögliche Koregulation der Zielgene durch den Transkriptionsfaktor hin.

Im Fall von ABI4 war bei einer Distanz von 96 bp relativ zum Beginn der ersten 5'UTR eine Koexpression der putativen Zielgene und des Faktors im Stängel zu verzeichnen. Die putativen Zielgene sind At3g11540, At4g15760, At4g34110, At4g39920, At4g15760, wobei At4g15760 jedoch bei Genevestigator nicht vertreten war (Daten nicht gezeigt). Bei der Distanz von 32 bp relativ zum Start der ersten 5'UTR konnten bei ABI4 und den putativen Zielgenen ebenfalls Gemeinsamkeiten bei der Regulation verzeichnet werden. Diese trat

ebenfalls bei den putativen Zielgenen und dem Transkriptionsfaktor im Stängel auf. Die putativen Zielgene haben die AGIs At1g60850, At2g34520, At2g48100, At3g10530, At3g55510, welche alle in Genevestigator vertreten sind. Diese Gene stimmen in keinem Fall mit denen, identifiziert für die Distanz 96 bp, überein.

Bei den putativen Zielgenen des Faktors TBP (Distanz 92) wurden Übereinstimmungen in der Genexpression festgestellt. Beim Transkriptionsfaktor TBP wird davon ausgegangen, dass er ubiquitär exprimiert wird. Weitere Gemeinsamkeiten zwischen der Koexpression der putativen Zielgene und dem jeweiligen Transkriptionsfaktor in der ersten 5'UTR konnten bislang nicht ermittelt werden. Im Anhang ist die Auswertung der Expressionsdaten detailliert aufgelistet (Tabelle 45).

Tabelle 19: Auswahl der 10 signifikantesten P-Werte in der ersten 5'UTR bei den pflanzlichen Matrizen. Die Distanzangabe bezieht sich auf den Abstand zwischen den bei diesen Distanzen vorkommenden putativen Transkriptionsfaktor-Bindungsstellen und dem Beginn dieses Bereichs. Die zu den putativen TFBSn gehörenden Gene (in dem Fall liegen diese innerhalb der Gene) sind numerisch aufgeführt. Die gemeinsame Regulation fasst Übereinstimmungen in der Genexpression bezüglich der Anatomie (A), der Entwicklung (D) sowie des zugeführten Stimulus (S) bei den bei Genevestigator eingesetzten putativen Zielgenen zusammen.

TF	Distanz [bp]	P-Wert	Anzahl Gene	Fehlende Gene bei Genevestigator	Gemeinsame Regulation
AGP1	52	0,00016447	6	0	D, A
ABI4(1)	96	0,0012739	4	1	D, A
ABI4(1)	44	0,00141236	5	0	D, A
ABI4(1)	32	0,0020814	5	0	D, A, S
ABI4(1)	26	0,0024586	5	0	D, A
ANT	79	0,00288803	4	0	D, A
AtMYB77	70	0,00049973	5	1	D, A, S
ID1	7	0,0006282	6	2	D, A, S
ID1	2	0,00067039	6	2	D, A
TBP	92	0,00174503	4	0	D, A, S

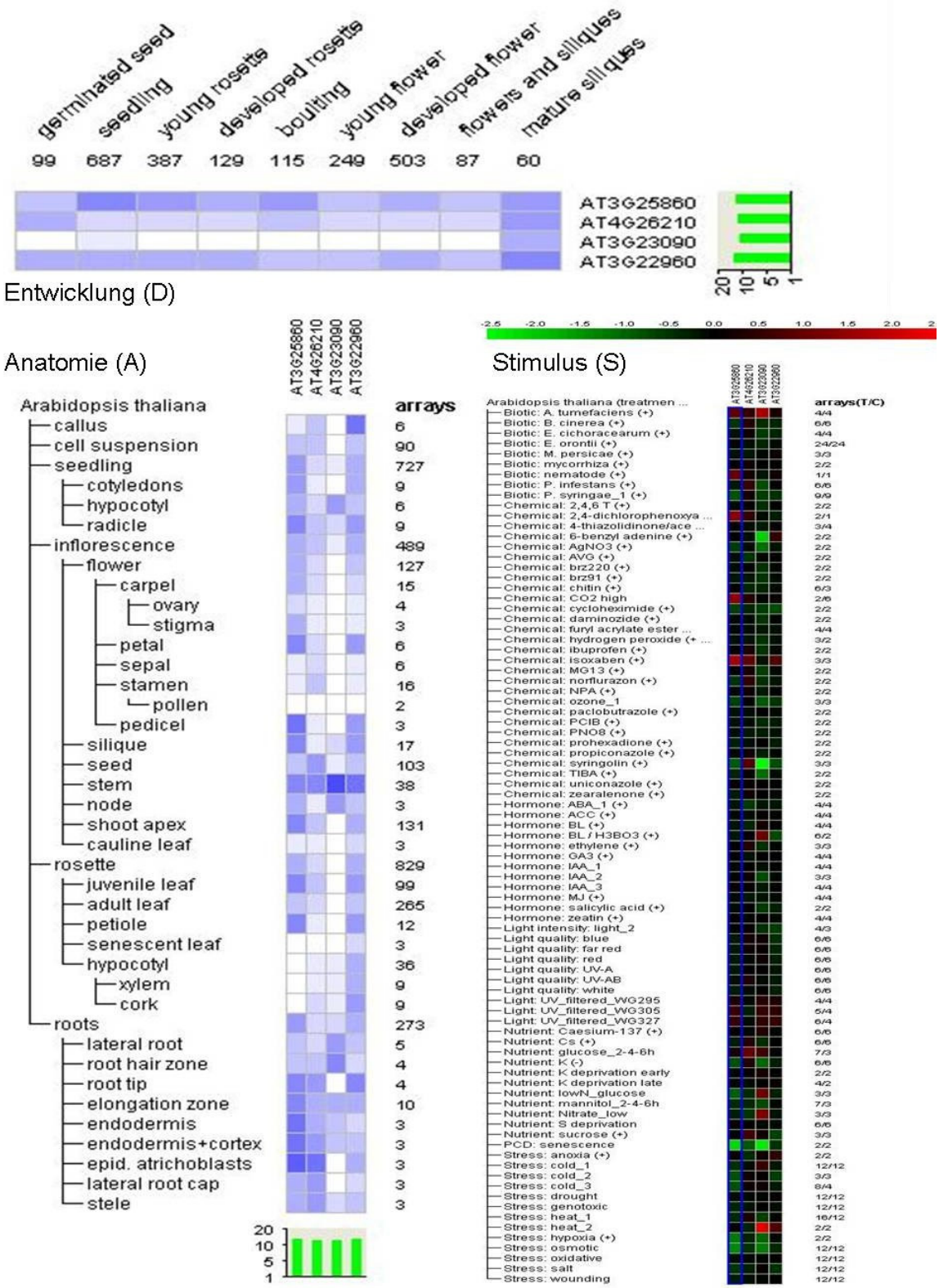


Abb. 51: Genevestigator-Grafiken der Expression der putativen Zielgene von AtMYB77, welcher putative Bindungsstellen in der ersten 5'UTR der Gene aufwies (Tabelle 19), die Erläuterungen sind im Text aufgeführt.

Diese Analysen, detailliert für die erste 5'UTR dargestellt, wurden auch für die anderen Bereiche durchgeführt, jedoch werden diese im Folgenden zusammenfassender aufgeführt.

Im ersten Intron konnten bei den 10 signifikantesten P-Werten in zwei Fällen Gemeinsamkeiten identifiziert werden. Bei den Transkriptionsfaktoren O2 und TGA1a (Tabelle 20) kam es bei beiden bei einer Distanz von 65 bp vom Beginn des ersten Introns aus betrachtet zu einer signifikanten Anreicherung der putativen Bindungsstellen. Die jeweils 16 putativen Zielgene sind rezeptorähnliche Protein-Kinasen, die auf dem vierten Chromosom fortlaufende AGIs aufweisen. Lediglich ein Gen ist auf dem Chromosom 3 lokalisiert. Bei Expressionsanalysen bei Genevestigator konnten drei Gene nicht aufgefunden werden, die restlichen wurden zusammengefasst dargestellt, was auf eine hochgradige Ähnlichkeit der Gene hindeutet. Im Fall des Transkriptionsfaktors TBP konnte bei den Zielgenen eine gemeinsame Regulation in ausgereiften Schoten und im Stängel identifiziert werden. Der Faktor wird ubiquitär exprimiert, so dass dieses Auftreten der übereinstimmenden Expressionsdaten nicht aussagekräftig ist (Daten nicht gezeigt). Ähnlichkeiten bei den restlichen sieben höchsten P-Werten konnten nicht identifiziert werden, es kam zu keiner einheitlichen Expression der putativen Zielgene. Die ausführlichen Ergebnisse gefinden sich im Anhang (Tabelle 46).

Tabelle 20: Auswahl der 10 signifikantesten P-Werte im ersten Intron bei den pflanzlichen Matrizen. Die Distanzangabe bezieht sich auf den Abstand zwischen den bei diesen Distanzen vorkommenden putativen Transkriptionsfaktor-Bindungsstellen und dem Beginn dieses Bereichs. Die zu den putativen TFBSn gehörenden Gene (in dem Fall liegen diese innerhalb der Gene) sind numerisch aufgeführt. Die gemeinsame Regulation fasst Übereinstimmungen in der Genexpression bezüglich der Anatomie (A), der Entwicklung (D) sowie des zugeführten Stimulus (S) bei den bei Genevestigator eingesetzten putativen Zielgenen zusammen.

TF	**Distanz**	**P-Wert**	**Anzahl Gene**	**Fehlende Gene bei Genevestigator**	**Gemeinsame Regulation**
AG_M00151	4	2,81E-27	28	1	-
AGL15	3	7,99E-21	23	0	-
ATHB9_M00417	14	1,19E-13	17	1	-
GT1	6	2,78E-64	53	12	-
MYBPH3_M00219	0	1,73E-10	14	2	-
O2	65	1,03E-12	16	3	Alle Gene zusammengefasst,

					Rezeptor ähnliche Kinase, kons. Nukl.
P	5	6,04E-20	23	2	-
RAV1_M00344	4	4,92E-61	51	4	-
TATABP	10	2,35E-09	13	1	D, A
TGA1a	65	1,12E-12	16	3	Alle Gene zusammengefasst, Rezeptor ähnliche Kinase, kons. Nukl.

Bei den zu den signifikantesten P-Werten des zweiten Introns gehörenden Daten konnten keinerlei Gemeinsamkeiten zwischen der Expression der putativen Zielgene und dem jeweiligen Transkriptionsfaktor ermittelt werden (Daten nicht gezeigt).

Die Ergebnisse für das dritte Intron sind in Tabelle 21 aufgeführt. Bei den Daten bezüglich des dritten Introns konnte in einem Fall eine mögliche Koregulierung des Transkriptionsfaktors AtMYB84 und der putativen Zielgene, At1g21480, At1g40133, At1g51820, At1g52150, At1g52670, At1g53510, At1g78060, At2g07310, At2g31840, At3g43940, At3g50240, At5g02320, At5g20850, At5g49950, At5g51220, At5g55040, innerhalb derer jeweils eine putative Transkriptionsfaktor-Bindungsstelle an der Position 7 relativ zum Start des dritten Introns lokalisiert war, festgestellt werden. Sowohl die putativen Zielgene als auch der Transkriptionsfaktor zeigen eine Expression in ausgereiften Schoten. Die Gene At1g40133 und At2g07310 waren jedoch in Genevestigator nicht vertreten.

Weitere Gemeinsamkeiten in Bezug auf eine Koregulierung konnten bei den anderen neun ausgewählten signifikanten P-Werten und den dazugehörigen Daten nicht ermittelt werden. Die ausführliche Tabelle im Anhang fasst die Ergebnisse zusammen (Tabelle 48).

Tabelle 21: Auswahl der 10 signifikantesten P-Werte im dritten Intron bei den pflanzlichen Matrizen. Die Distanzangabe bezieht sich auf den Abstand zwischen den bei diesen Distanzen vorkommenden putativen TFBSn und dem Beginn dieses Bereichs. Die zu den putativen TFBSn gehörenden Gene, in dem Fall liegen diese innerhalb der Gene, sind numerisch aufgeführt. Die gemeinsame Regulation fasst Übereinstimmungen in der Genexpression bezüglich der Anatomie (A), der Entwicklung (D) sowie des zugeführten Stimulus (S) bei den bei Genevestigator eingesetzten putativen Zielgenen zusammen.

TF	Distanz	P-Wert	Anzahl Gene	Fehlende Gene bei Genevestigator	Gemeinsame Regulation
AG_M00151	4	1,28E-27	25	0	-
AGL15	3	9,06E-13	14	2	A
AtMYB15	7	4,99E-14	15	0	-
AtMYB84	7	3,27E-15	16	2	D
bZIP911_2	0	6,75E-06	6	1	D, A
GT1	6	7,58E-58	44	4	-
MYBPH3_M00219	0	3,78E-05	7	1	D, A
P	5	1,62E-29	27	2	-
RAV1_M00344	4	4,48E-51	40	2	-
TEIL_M00502	5	1,10E-99	69	6	-

Im vierten Intron (Tabelle 22) konnten bei den putativen Zielgenen von AtMYB84, welcher an der Distanz 7 relativ zum Beginn des Introns putative TFBSn aufwies, Gemeinsamkeiten in der Genexpression verzeichnet werden, jedoch stimmten diese nicht mit den Expressionsdaten von AtMYB84 überein. Ähnliches konnte für MYB.PH3 an der Distanz 52 festgestellt werden. Insgesamt kam es nicht zu übereinstimmenden Genexpressionen. Im Anhang sind die entsprechenden Daten in Tabelle 49 angefügt.

Tabelle 22: Auswahl der 10 signifikantesten P-Werte im vierten Intron bei den pflanzlichen Matrizen. Die Distanzangabe bezieht sich auf den Abstand zwischen den bei diesen Distanzen vorkommenden putativen TFBSn und dem Beginn dieses Bereichs. Die zu den putativen TFBSn gehörenden Gene, in dem Fall liegen diese innerhalb der Gene, sind numerisch aufgeführt. Die gemeinsame Regulation fasst Übereinstimmungen in der Genexpression bezüglich der Anatomie (A), der Entwicklung (D) sowie des zugeführten Stimulus (S) bei den bei Genevestigator eingesetzten putativen Zielgenen zusammen.

TF	Distanz	P-Wert	Anzahl Gene	Fehlende Gene bei Genevestigator	Gemeinsame Regulation

AG_M00151	4	2,90E-31	26	4	-
AGL15	3	1,42E-16	16	1	-
AGL3	4	3,91E-10	11	2	-
AtMYB15	7	6,15E-14	14	0	-
AtMYB84	7	1,38E-12	13	0	D, A
GT1	6	1,25E-26	23	1	-
MYBPH3_M00219	52	1,04E-05	7	1	S
P	5	9,56E-26	23	2	-
RAV1_M00344	4	3,48E-44	34	3	-
TEIL_M00502	5	8,55E-100	66	4	-

Im Fall von AGL15, welcher putative Bindungsstellen an der Distanz 3 des fünften Introns (Tabelle 23) bei den entsprechenden putativen Zielgenen aufwies, wurde eine übereinstimmende Genexpression der Gene und des Transkriptionsfaktors in ausgereiften Schoten ermittelt. Die putativen Zielgene des Faktors, At1g29330, At1g63780, At3g01510, At5g01270, At5g59740 und At5g65420, waren alle bei Genevestigator vorhanden. Die Daten befinden sich im Anhang (Tabelle 50).

Bei den putativen Zielgenen des Faktors AtMYB84, welcher putative Bindungsstellen an der Distanz 7 aufwies, konnte eine übereinstimmende Genexpression der Gene und des Transkriptionsfaktors in ausgereiften Schoten, im Hypocotyl und in der Blüte festgestellt werden. Expressionsdaten bezüglich der putativen Zielgene At1g12520, At1g20570, At1g38131, At1g66530, At2g17900, At5g16650, At5g23300 sowie At5g40270 konnten mittels Genevestigator belegt werden (Tabelle 50 im Anhang).

Des Weiteren wurde eine Koexpression der putativen Zielgene von AG (Distanz 4), AtMYB15 (Distanz 7), MYBPH3 (Distanzen 0 und 12) bzw. RAV1 (Distanz 4) ermittelt, jedoch stimmten diese nicht mit den Expressionsdaten bezüglich des jeweiligen Transkriptionsfaktors überein (Tabelle 50 im Anhang). Weitere Übereinstimmungen im fünften Intron konnten nicht aufgezeigt werden.

Tabelle 23: Auswahl der 10 signifikantesten P-Werte im fünften Intron bei den pflanzlichen Matrizen. Die Distanzangabe bezieht sich auf den Abstand zwischen den bei diesen Distanzen vorkommenden putativen Transkriptionsfaktor-Bindungsstellen und dem Beginn dieses Bereichs. Die zu den putativen TFBSn gehörenden Gene, in dem Fall liegen diese innerhalb der Gene, sind numerisch aufgeführt. Die gemeinsame Regulation fasst Übereinstimmungen in der Genexpression bezüglich der Anatomie

(A), der Entwicklung (D) sowie des zugeführten Stimulus (S) bei den bei Genevestigator eingesetzten putativen Zielgenen zusammen.

TF	Distanz	P-Wert	Anzahl Gene	Fehlende Gene bei Genevestigator	Gemeinsame Regulation
AG_M00151	4	2,18E-07	8	2	D, A
AGL15	3	3,57E-05	6	0	D, A
AtMYB15	7	2,11E-07	8	0	D, A
AtMYB84	7	2,30E-07	8	1	D, A, S
GT1	6	7,61E-37	28	2	-
MYBPH3_M00219	0	1,36E-08	9	0	D, A
MYBPH3_M00219	12	3,46E-05	6	0	D
P	5	6,75E-31	25	2	A
RAV1_M00344	4	4,69E-11	11	2	D, A
TEIL_M00502	5	1,06E-59	42	3	-

Die Ergebnisse für das erste Exon sind in Tabelle 24 aufgeführt. Bei den ausgewählten 10 signifikantesten P-Werten im ersten Exon bei den zu den putativen Transkriptionsfaktor-Bindungsstellen zugeordneten Genen des Faktors ABI4 (Distanzen 41, 74 sowie 50) konnten jeweils Gemeinsamkeiten bezüglich der Genexpression der jeweiligen putativen Zielgene festgestellt werden (Tabelle 51 im Anhang). Jedoch konnte keine Korrelation mit den Genexpressionsdaten des Transkriptionsfaktors identifiziert werden. Die zu dem Transkriptionsfaktor zugehörigen putativen Bindungsstellen stellten sieben der 10 signifikantesten P-Werte dar. An den restlichen vier Distanzen konnte jedoch keinerlei gemeinsame Koexpression der putativen Zielgene festgestellt werden (Tabelle 51 im Anhang).

Die putativen Bindungsstellen von zwei weiteren Transkriptionsfaktoren waren im ersten Exon von putativen Zielgenen lokalisiert, welche jeweils Gemeinsamkeiten aufwiesen. Bei den zu bZIP910 zugeordneten Genen, innerhalb derer der Faktor an der Position 112 relativ vom Beginn des Exons aus betrachtet putative Bindungsstellen aufwies, trat eine Expression der putativen Zielgene in reifen Schoten auf. Es konnte jedoch keine Übereinstimmung mit der Expression des Transkriptionsfaktors verzeichnet werden (Tabelle 51 im Anhang).

Im Fall von RAV1 (Distanz 464) kam es bei den putativen Zielgenen zu einer gemeinsamen Expression in ausgereiften Schoten sowie zu einer Herunterregulation unter Isoxaben-Einfluss und einer Hochregulation bei zu wenig Nitrat-Zugabe. Jedoch konnte auch hier

keine Übereinstimmung dieser Expressionsdaten mit denen für den Transkriptionsfaktor festgestellt werden (Tabelle 51 im Anhang).

Tabelle 24: Auswahl der 10 signifikantesten P-Werte im ersten Exon bei den pflanzlichen Matrizen. Die Distanzangabe bezieht sich auf den Abstand zwischen den bei diesen Distanzen vorkommenden putativen Transkriptionsfaktor-Bindungsstellen und dem Beginn dieses Bereichs. Die zu den putativen TFBSn gehörenden Gene, in dem Fall liegen diese innerhalb der Gene, sind numerisch aufgeführt. Die gemeinsame Regulation fasst Übereinstimmungen in der Genexpression bezüglich der Anatomie (A), der Entwicklung (D) sowie des zugeführten Stimulus (S) bei den bei Genevestigator eingesetzten putativen Zielgenen zusammen.

TF	Distanz	P-Wert	Anzahl Gene	Fehlende Gene bei Genevestigator	Gemeinsame Regulation
ABI4_1	29	2,56E-26	30	4	-
ABI4_1	41	2,58E-25	29	1	D
ABI4_1	53	2,41E-24	28	3	-
ABI4_1	74	1,10E-23	27	3	D, A
ABI4_1	62	3,70E-22	26	2	-
ABI4_1	50	1,09E-20	25	2	D
ABI4_1	104	1,11E-19	23	3	-
ATHB5	7	6,91E-32	35	2	-
bZIP910_2	112	1,38E-19	23	4	D (kons. Nukl.), Retrotransposons
RAV1_M00344	464	3,68E-28	24	14	D, S (kons. Nukl.), Retrotransposons

Im zweiten Exon (Tabelle 25) zeigten die putativen Zielgene (At3g04180, At3g04190, At3g18810, At3g63410, At5g39130, At5g39160 und At5g39190) des Transkriptionsfaktors PIF3 (Distanz 227) eine gemeinsame Expression in ausgereiften Schoten, im Kallus, in der Keimwurzel und in der Wurzel. In den Schoten war keine Expression nachweisbar. Dies stimmt mit den Expressionsdaten für den Transkriptionsfaktor überein (Tabelle 52 im Anhang). Teilweise waren die Gene bei Genevestigator zusammengefasst, was auf Genfamilien hindeutet. Der Großteil gehört zu den Germin-ähnlichen Proteinen.

Im Fall von RAV1 (Distanz 4) konnte eine gemeinsame Expression der putativen Zielgene in ausgereiften Schoten und im Stängel detektiert werden, welche sich jedoch nicht mit der Expression des Faktors deckte. Ähnliches galt für die putativen Zielgene der Transkriptionsfaktoren P (Distanz 5) sowie AGP1 (Distanz 5), wobei es hier bei den entsprechenden Genen zu einer Expression kam, welche beim Transkriptionsfaktor nicht zu verzeichnen war. Bei den putativen Zielgenen der Transkriptionsfaktoren ID1 und bZIP910 handelte es sich vermutlich um stark konservierte Genfamilien, da die entsprechenden Gene jeweils nur zusammengefasst in der Datenbank vertreten waren (Tabelle 52 im Anhang).

Tabelle 25: Auswahl der 10 signifikantesten P-Werte im zweiten Exon bei den pflanzlichen Matrizen. Die Distanzangabe bezieht sich auf den Abstand zwischen den bei diesen Distanzen vorkommenden putativen Transkriptionsfaktor-Bindungsstellen und dem Beginn dieses Bereichs. Die zu den putativen TFBSn gehörenden Gene, in dem Fall liegen diese innerhalb der Gene, sind numerisch aufgeführt. Die gemeinsame Regulation fasst Übereinstimmungen in der Genexpression bezüglich der Anatomie (A), der Entwicklung (D) sowie des zugeführten Stimulus (S) bei den bei Genevestigator eingesetzten putativen Zielgenen zusammen.

TF	Distanz	P-Wert	Anzahl Gene	Fehlende Gene bei Genevestigator	Gemeinsame Regulation
AGP1	5	9,89E-14	17	1	D
bZIP910_2	162	4,02E-16	15	3	Alle Gene zusammengefasst, Rezeptor ähnliche Kinase, kons. Nukl.
HVH21	1	1,15E-06	10	2	A, S
ID1	258	4,06E-19	15	3	Alle Gene zusammengefasst, Rezeptor ähnliche Kinase, kons. Nukl.
MYBPH3_M00218	196	4,61E-09	9	1	-
P	5	4,34E-09	13	1	D
PIF3_2	227	5,04E-07	7	0	D, A, S, Germin ähnliche Proteine, (kons. Nukl.)
RAV1_M00344	4	1,32E-41	38	6	D, A

TATABP	0	4,61E-26	27	3	-
TEIL_M00502	5	2,89E-34	34	3	-

Nach der Auswahl der 10 signfikantesten P-Werte im dritten Exon (Tabelle 26) und der Analyse der putativen Transkriptionsfaktor-Bindungsstellen sowie der jeweiligen putativen Zielgene (At1g24350, At1g80390, At2g03390, At2g46410, At3g29190, At4g23060, At5g07800, At5g55250, At5g55290 und At5g61070) konnte bei denen des Transkriptionsfaktors AtMYB84 (Distanz 3) eine gemeinsame Expression in reifen Schoten festgestellt werden, was sich mit der Expression des Transkriptionsfaktors deckte (Tabelle 53 im Anhang). At1g80390 war bei Genevestigator jedoch nicht vorhanden.

Die putativen Zielgene der Faktoren AGP1 (Distanz 5), AtMYB15 (Distanz 3), bZIP910_2 (Distanz 43) und P (Distanz 5) zeigten eine gemeinsame Expression, jedoch konnten diese nicht mit den Expressionsdaten bezüglich des jeweiligen Transkriptionsfaktors in Verbindung gebracht werden. Weitere Gemeinsamkeiten konnten nicht beobachtet werden (Tabelle 53 im Anhang).

Tabelle 26: Auswahl der 10 signifikantesten P-Werte im dritten Exon bei den pflanzlichen Matrizen. Die Distanzangabe bezieht sich auf den Abstand zwischen den bei diesen Distanzen vorkommenden putativen Transkriptionsfaktor-Bindungsstellen und dem Beginn dieses Bereichs. Die zu den putativen TFBSn gehörenden Gene, in dem Fall liegen diese innerhalb der Gene, sind numerisch aufgeführt. Die gemeinsame Regulation fasst Übereinstimmungen in der Genexpression bezüglich der Anatomie (A), der Entwicklung (D) sowie des zugeführten Stimulus (S) bei den bei Genevestigator eingesetzten putativen Zielgenen zusammen.

TF	**Distanz**	**P-Wert**	**Anzahl Gene**	**Fehlende Gene bei Genevestigator**	**Gemeinsame Regulation**
ABI4_2	32	1,57E-06	9	1	-
AG_M00151	4	1,72E-06	9	1	-
AGP1	5	2,17E-07	10	1	D, A
AtMYB15	3	1,66E-06	9	1	D, A
AtMYB84	3	1,91E-07	10	1	D
bZIP910_2	43	1,24E-05	8	1	D, A, S
P	5	2,86E-05	8	0	D, A
RAV1_M00344	4	9,61E-26	25	0	-
TATABP	0	6,61E-23	23	2	-
TEIL_M00502	5	2,60E-27	27	3	-

Im vierten Exon (Tabelle 27) kam es im Fall der putativen Zielgene (At1g17930, At1g30330, At2g04842, At2g23470, At3g53780, At4g04570 und At5g23080) des Transkriptionsfaktors AG, wobei die putativen Bindungsstellen des Faktors eine Distanz von 4 Nukleotiden relativ vom Beginn des Exons aus betrachtet aufwiesen, zu gemeinsamen Expressionen in der Blüte und in der Schote. Diese stimmten mit den Expressionsdaten des Transkriptionsfaktors überein (Tabelle 54 im Anhang). Für At2g04842 standen keine Expressionsdaten zur Verfügung.

Die putativen Zielgene der Faktoren AGP1 (Distanz 5), AtMYB15 (Distanzen 3 und 71), HVH21 (Distanz 1), RAV1 (Distanz 4) sowie TBP (Distanz 0) wiesen eine gemeinsame Expression auf, die jedoch bis auf die Expression des TBP nicht mit der Expression des jeweiligen Transkriptionsfaktors korreliert werden konnte. Bei TBP wird davon ausgegangen, dass der Faktor ubiquitär exprimiert wird. Bei den zu GAMYB zugeordneten Genen, bei denen der Faktor eine putative Bindungsstelle an der Position 183 relativ zum Beginn des Exons besaß, handelte es sich bei diesen Genen vermutlich um eine hoch konservierte Genfamilie, da die Gene bei Genevestigator zusammengefasst dargestellt wurden. Die restlichen ausgewählten Signifikanzen zeigten keinerlei Gemeinsamkeiten (Tabelle 54 im Anhang).

Tabelle 27: Auswahl der 10 signifikantesten P-Werte im vierten Exon bei den pflanzlichen Matrizen. Die Distanzangabe bezieht sich auf den Abstand zwischen den bei diesen Distanzen vorkommenden putativen Transkriptionsfaktor-Bindungsstellen und dem Beginn dieses Bereichs. Die zu den putativen TFBSn gehörenden Gene, in dem Fall liegen diese innerhalb der Gene, sind numerisch aufgeführt. Die gemeinsame Regulation fasst Übereinstimmungen in der Genexpression bezüglich der Anatomie (A), der Entwicklung (D) sowie des zugeführten Stimulus (S) bei den bei Genevestigator eingesetzten putativen Zielgenen zusammen.

TF	Distanz	P-Wert	Anzahl Gene	Fehlende Gene bei Genevestigator	Gemeinsame Regulation
AG_M00151	4	3,25E-05	7	1	D, A
AGP1	5	3,72E-05	7	0	D, A, S
AtMYB15	3	3,16E-05	7	0	D, A
AtMYB15	71	9,66E-05	6	0	D, A
AtMYB77	70	0,00010261	6	0	-
GAMYB	183	5,73E-18	13	1	Alle Gene zusammengefasst, AGIs

					aufeinanderfolgend, Rezeptor ähnliche Kinasen, kons. Nukl.
HVH21	1	3,22E-05	7	1	D, A
RAV1_M00344	4	1,52E-23	15	4	D
TATABP	0	2,39E-15	16	3	D
TEIL_M00502	5	3,54E-21	21	0	-

Bei den zu den 10 signifikantesten P-Werten gehörenden Daten in der ersten 3'UTR (Tabelle 28) kam es in drei Fällen, bei den Transkriptionsfaktoren ATHB5 (Distanzen 6 und 208) sowie CDC5 (Distanz 2), jeweils zu einer gemeinsamen Regulation der zu den putativen TFBSn zugeordneten Genen. Die putativen Zielgene von ATHB5, bei denen putative TFBSn des Faktors bei einer Distanz von 6 Nukleotiden relativ zum Beginn der 3'UTR lokalisiert waren, zeigten eine gemeinsame Expression in der Endodermis, wobei ATHB5 ebenfalls eine Expression in dem Gewebe aufwies. Die entsprechenden Gennamen sind At1g01120, At1g67590, At2g02000, At2g14170, At2g36820, At5g14690 sowie At5g24970. Bei den zu der Signifikanz an der Distanz 208 gehörenden putativen Zielgenen wurde eine Herunterregulation der Genexpression durch Zugabe von Syringolin verzeichnet, was sich mit den Expressionsdaten des Transkriptionsfaktors deckte. Die Gene sind At1g03600, At3g07990, At3g43860 und At3g60080. Bei den zu dem Transkriptionsfaktor CDC5 zugehörigen Genen (At1g70710, At1g76650, At1g76970, At1g79750, At2g36900, At4g29050 und At4g32130) konnte sowohl bei den Genen als auch bei dem Faktor eine Expression in ausgereiften Schoten sowie im Stängel detektiert werden (Tabelle 55 im Anhang).

Die putativen Zielgene des Faktors ATHB1 wiesen bei den signifikanten Distanzen 99, 89 und 2 jeweils gemeinsame Expressionsdaten auf, welche jedoch nicht mit denen des Transkriptionsfaktors übereinstimmten. Gleiches galt für die putativen Zielgene des Transkriptionsfaktors HAHB4 (Distanz 5) und die Expression des Faktors selbst (Tabelle 55 im Anhang). Weitere Gemeinsamkeiten konnten nicht identifiziert werden.

Tabelle 28: Auswahl der 10 signifikantesten P-Werte in der ersten 3'UTR bei den pflanzlichen Matrizen. Die Distanzangabe bezieht sich auf den Abstand zwischen den bei diesen Distanzen vorkommenden putativen Transkriptionsfaktor-Bindungsstellen und dem Beginn dieses Bereichs. Die zu den putativen TFBSn gehörenden Gene, in dem Fall liegen diese innerhalb der Gene, sind numerisch aufgeführt. Die gemeinsame Regulation fasst Übereinstimmungen in der Genexpression bezüglich der Anatomie (A), der Entwicklung (D) sowie des zugeführten Stimulus (S) bei den bei Genevestigator eingesetzten putativen Zielgenen zusammen.

TF	Distanz	P-Wert	Anzahl Gene	Fehlende Gene bei Genevestigator	Gemeinsame Regulation
ATHB1	99	0,00016804	7	0	D, A
ATHB1	89	0,00117754	6	1	D, A
ATHB1	2	0,00194389	6	0	S
ATHB5	6	0,00015627	7	0	A
ATHB5	208	0,0017943	4	0	S
ATHB9_M00417	10	0,00100632	6	0	-
CDC5_M00361	2	0,00018724	7	0	D, A
HAHB4	5	0,00196242	6	0	S
MYBPH3_M00218	8	2,48E-07	10	0	-
MYBPH3_M00219	8	0,00014517	7	0	-

Die putativen Bindungsstellen im intergenischen Bereich wurden auf verschiedene Weise analysiert. Die TFBSn, deren zugeordnete Gene einen Transkriptionsstartpunkt aufwiesen, wurden von denen, bei denen die Gene nur einen bekannten Translationsstartpunkt hatten, getrennt betrachtet.

Bei den signifikantesten 10 P-Werten und den dazugehörenden Daten bezüglich der putativen Transkriptionsfaktor-Bindungsstellen und der dazugehörenden Gene, unter der Bedingung, dass der Transkriptionsstartpunkt annotiert war (Tabelle 29), konnte bei den putativen Zielgenen von O2, bei einer Distanz von -79, das heißt upstream vom dazugehörenden Gen, eine gemeinsame Regulation in gekeimten Samen festgestellt werden. Der Transkriptionsfaktor wird im Endosperm exprimiert, welches das Nährgewebe für den keimenden Samen darstellt (Singh, 1998). Tabelle 56 im Anhang fasst die Ergebnisse zusammen. Die Gennamen sind At1g59900, At1g68650, At1g70480, At2g18240, At2g43330, At3g02230, At3g05730, At3g10330, At3g59770, At4g24800, At4g33950, At5g02120, At5g39660 und At5g58070.

Die zu O2 gehörenden Gene, bei denen die putativen TFBSn bei -92 und -88 liegen, konnte jeweils eine gemeinsame Expression der putativen Zielgene verzeichnet werden, jedoch war

diese nicht mit den Expressionsdaten bezüglich des Faktors deckungsgleich. Des Weiteren wurden bei den putativen Zielgenen von ABF1 (Distanz -78) übereinstimmende Expressionsdaten in ausgereiften Schoten festgestellt, was jedoch nicht mit der Expression des Faktors übereinstimmte. Weitere Ähnlichkeiten wurden in dem Bereich nicht festgestellt.

Tabelle 29: Auswahl der 10 signifikantesten P-Werte intergenischen Bereich bei den pflanzlichen Matrizen. Bedingung war, dass das zu der Bindungsstelle annotierte Gen einen Transkriptionsstart aufweist. Die Distanzangabe bezieht sich auf den Abstand zwischen den bei diesen Distanzen vorkommenden putativen Transkriptionsfaktor-Bindungsstellen und dem Beginn dieses Bereichs. Die zu den putativen TFBSn gehörenden Gene, in dem Fall liegen diese innerhalb der Gene, sind numerisch aufgeführt. Die gemeinsame Regulation fasst Übereinstimmungen in der Genexpression bezüglich der Anatomie (A), der Entwicklung (D) sowie des zugeführten Stimulus (S) bei den bei Genevestigator eingesetzten putativen Zielgenen zusammen.

TF	Distanz	P-Wert	Anzahl Gene	Fehlende Gene bei Genevestigator	Gemeinsame Regulation
ABF1	-78	9,64E-27	17	0	D
ABF1	-88	6,87E-25	17	3	-
O2	-92	6,33E-22	15	0	D
O2	-66	7,98E-22	15	0	-
O2	-88	3,87E-20	14	0	D, A
O2	-79	4,20E-20	14	0	D
ABF1	-85	2,46E-19	15	2	-
O2	-105	1,92E-18	14	3	-
O2	-72	2,47E-18	13	1	-
ABF1	-66	1,75E-17	13	1	-

Wurden bei der Analyse nur diejenigen putativen TFBSn eingesetzt, deren zugeordnetes Gen lediglich einen annotierten Translationsstartpunkt aufwies (Tabelle 30), kam es in drei Fällen zu Gemeinsamkeiten bezüglich der Expression der Gene und der des jeweiligen Transkriptionsfaktors (Tabelle 57 im Anhang). Die zu AtMYB15 (Distanz -56) gehörenden Gene wurden unter dem Einfluss von AVG (Aminoethoxyvinylglycin) herunterreguliert, was ebenfalls bei dem Transkriptionsfaktor zu verzeichnen war. Die 10 Gene gehören zu den Retrotransposons. Im Fall der zu RAV1 gehörenden Gene (Distanz -687) trat eine Herunterregulierung der putativen Zielgene durch die Zugabe von Propiconazol bzw. Uniconazol auf, während es bei Wärmestress zu einer Hochregulierung der putativen Zielgene kam. Diese Expressionsdaten stimmen mit denen von RAV1 überein. 7 der 8

putativen Zielgene gehören zu den Retrotransposons, eines codiert für einen MYB Transkriptionsfaktor. Die zu AGL15 gehörenden Gene (Distanz -8) wiesen unter Zugabe des Antiauxins PCIB (a-p-Chlorophenoxy-Isobuttersäure) eine Hochregulierung auf, während es bei der Expression des Transkriptionsfaktors zu einer Herunterregulierung unter demselben Einfluss kam. Die AGIs der Gene sind At1g60980, At1g62580, At2g29370, At3g09290, At3g51210, At3g52520, At4g34650 und At5g06460, wobei At3g52520 nicht bei Genevestigator vertreten war (Tabelle 57 im Anhang).

Die putativen Zielgene von MYB.PH3 zeigten eine gemeinsame Expression in der Zellsuspension, jedoch kam diese bei dem Transkriptionsfaktor nicht vor (Matys *et al.*, 2003).

Im Fall von HAHB4 (Distanz -232) wurde eine Expression der putativen Zielgene im Stängel detektiert. Ob diese bei dem Faktor ebenfalls auftritt, ist unbekannt. Der Faktor stammt aus der Sonnenblume und ist somit bei Genevestigator nicht vertreten, bei TRANSFAC© ist nichts über die Expression des Faktors angegeben (Matys *et al.*, 2003).

Die putativen Zielgene von GAMYB (Distanz -502) waren bei Genevestigator zusammengefasst, was auf eine hochkonservierte Genfamilie hindeutet (Tabelle 57 im Anhang). Weitere Gemeinsamkeiten wurden nicht identifiziert.

Tabelle 30: Auswahl der 10 signifikantesten P-Werte intergenischen Bereich bei den pflanzlichen Matrizen. Bedingung war, dass das zu der Bindungsstelle annotierte Gen nur einen Translationsstart aufweist. Die Distanzangabe bezieht sich auf den Abstand zwischen den bei diesen Distanzen vorkommenden putativen Transkriptionsfaktor-Bindungsstellen und dem Beginn dieses Bereichs. Die zu den putativen TFBSn gehörenden Gene, in dem Fall liegen diese innerhalb der Gene, sind numerisch aufgeführt. Die gemeinsame Regulation fasst Übereinstimmungen in der Genexpression bezüglich der Anatomie (A), der Entwicklung (D) sowie des zugeführten Stimulus (S) bei den bei Genevestigator eingesetzten putativen Zielgenen zusammen.

TF	Distanz	P-Wert	Anzahl Gene	Fehlende Gene bei Genevestigator	Gemeinsame Regulation
AGL15	-8	4,22E-08	8	1	D, S
ATHB5	-4	2,35E-16	14	3	-
AtMYB15	-56	9,78E-11	10	0	D, A, S, Retrotransposon, (kons. Nukl.)
GAMYB	-502	1,24E-08	8	7	Nicht repräsentativ (1 von 8 Gene)
HAHB4	-232	2,42E-10	10	2	A

MYBPH3_M00218	-85	3,85E-08	8	1	A, S
MYBPH3_M00218	-204	5,28E-07	7	5	Nicht repräsentativ (2 von 5 Gene)
RAV1_M00343	-13	4,81E-12	11	8	Nicht repräsentativ (3 von 11 Gene)
RAV1_M00343	-9	1,12E-10	10	9	Nicht repräsentativ (1 von 10 Gene)
RAV1_M00343	-687	6,68E-09	8	2	S

Zusammengefasst konnte unter Verwendung des Signifikanzkriteriums und durch Auswahl der 10 signfikantesten P-Werte gezeigt werden, dass es in einigen Fällen eine gemeinsame Expression der putativen Zielgene und dem Transkriptionsfaktor gibt, was darauf hindeutet, dass sich weitere Analysen, diesmal jedoch im Labor, anbieten würden, um die tatsächliche gemeinsame Regulierung durch Bindung des jeweiligen Transkriptionsfaktors belegen zu können. Insgesamt kam es nicht nur im intergenischen Bereich zu möglichen Koregulationen, sondern auch in Bereichen innerhalb von Genen, insbesondere in der 5'UTR und der 3'UTR.

4. Diskussion

4.1 Besonderheiten bei den Genen und der Nukleotidkomposition von *Arabidopsis thaliana*

Zu Beginn dieser Arbeit lag der *Arabidopsis thaliana* Release 5 vor (Haas *et al*., 2005). Die Daten bezüglich der Genannotation sowie der DNA-Sequenzen waren in Form von XML-Dateien verfügbar. Eingesetzt wurden für diese Arbeit die XML-Dateien für die fünf *Arabidopsis thaliana* Chromosomen. Diese wurden mit Perl-Skripten geparsed und in die bestehende Datenbankstruktur der TFBSites implementiert. Verwendet wurden nur die proteinkodierenden Gene, zu denen in den XML-Dateien ebenfalls die Transposons sowie Pseudogene zählten. Des Weiteren waren Splicevarianten der Gene ebenfalls annotiert. Aufgrund der durchzuführenden Verteilungsanalysen von Transkriptionsfaktor-Bindungsstellen wurde das Genom zunächst analysiert.

Splicevarianten wurden aus den Analysen ausgeschlossen, dadurch kam es zu einer Gesamtanzahl an proteinkodierenden Genen von 29993. Der intergenische Bereich ist im Schnitt 1793 Nukleotide lang, was bei einer Gesamtlänge von 119186497 Nukleotiden und einer durchschnittlichen Länge der Gene von 2246 Nukleotiden zu einer Lokalisierung von einem Gen pro 4000 Nukleotide führt. Das Genom von *Arabidopsis* ist damit recht dicht gepackt.

Gene bestehen in der Regel aus einer 5'UTR, aus Exons, Introns und einer 3'UTR. Introns können nicht nur zwischen Exons lokalisiert sein sondern können auch innerhalb der UTRs liegen. Um einen Eindruck zu erhalten, wie viele dieser Bereiche sich durchschnittlich in einem Gen befinden und wie lang sie jeweils sind, wurden diese bestimmt. Ausgehend vom Transkriptionsstartpunkt wurden die reinen UTRs, also ohne Introns, analog zu den anderen Bereichen eines Gens durchnummeriert. Es wurde festgestellt, dass die meisten Gene UTRs ohne Introns aufweisen. Die durchschnittliche Anzahl der 5'UTR beträgt 1,19, die der 3'UTR 1,09 pro Gen. Diese Zahlenwerte kommen z.B. dadurch zustande, dass eine 5'UTR mit einem enthaltenen Intron als zwei 5'UTRs gezählt wurde. Im Fall der Exons wurde eine durchschnittliche Anzahl von 4,72 ermittelt, was mit der Angabe von 5,42 (Haas *et al*., 2005) nicht ganz übereinstimmt. Der Unterschied resultiert vermutlich darin, dass einmal die Auswahl bestimmter Gene limitiert wurde, andererseits alle Exons in die Analyse mit eingingen. Dennoch stimmen die Mittelwerte gut überein. Die durchschnittliche Anzahl der Introns pro Gen beträgt 5,48. Es konnte gezeigt werden, dass Introns für eine effiziente Pol II Transkription essentiell sind (Kornblihtt *et al.,* 2004). Aber auch die Anzahl der Introns in

einem Gen kann von Bedeutung sein. Bei Genen mit weniger als vier Introns ist das durchschnittliche mRNA Level herabgesetzt (Alexandrov *et al.,* 2006).
Neben der Anzahl der einzelnen Bereiche innerhalb der Gene wurde die durchschnittliche Länge ebenfalls bestimmt. Diese beträgt für die erste 5'UTR 119 Nukleotide. Die 5'UTR in der Hefe ist in der Regel 20-60 bp lang, im Mittel 52 bp (Day und Tuite, 1998). Jedoch ist die 5'UTR nicht essentiell. Am 5' Ende der mRNA befindet sich bei Eukaryonten die 5' Cap-Struktur, die als m^7G abgekürzt wird, wobei das Guanin methyliert vorliegt. Diese Struktur schützt die mRNAs vor dem 5'-3' exonuklearen Abbau und stimuliert die Translationseffizienz. Die mRNA des Hefegens TCM1 weist beispielsweise keine 5'UTR auf, besitzt aber eine Cap-Struktur an der Position +1, und kann translatiert werden, wenn auch ineffizient (Day und Tuite, 1998). In *Arabidopsis thaliana* inhibiert eine sehr kurze 5'UTR (<20 Nukleotide) den Eintritt des 43S Prä-Initiationskomplexes oder die Erkennung des AUG Initiationscodons für die Translation. Weist die 5'UTR eine angemessene Länge von 40 bis 100 Nukleotiden auf, wird die Initiation der Translation begünstigt (Kawaguchi und Bailey-Serres, 2005). Dies stimmt in etwa mit der in dieser Arbeit ermittelten durchschnittlichen Länge der 5'UTR überein.
Die durchschnittliche Länge der Exons beträgt laut Haas *et al*. (2005) 276 Nukleotide, die der Introns 164 Nukleotide. In der vorliegenden Arbeit wurde die durchschnittliche Länge detaillierter analysiert, die ersten Exons (692 Nukleotide) sind weitaus länger als alle Exons im Durchschnitt. Die darauffolgenden Exons sind dann entsprechend kürzer. Gleiches gilt für die Introns, das erste ist länger (248 Nukleotide) als die folgenden. Bei beiden Bereichen nimmt die Anzahl der durchschnittlichen Nukleotide mit zunehmender Anzahl innerhalb eines Gens ab.
Die 3'UTR ist im Durchschnitt 223 Nukleotide lang. Der Effekt der Länge und anderer Merkmale der 3'UTR auf die Translation wurde bislang in Pflanzen nicht umfassend untersucht (Kawaguchi und Bailey-Serres, 2005).

Neben der Anzahl der einzelnen Bereiche innerhalb von Genen wurde der AT-Gehalt (respektive GC-Gehalt) des *Arabidopsis thaliana* Genoms ermittelt. Dieser gilt als taxonomisches Merkmal, da er auch zwischen verschiedenen Arten unterschiedlich ist. Der durchschnittliche AT-Gehalt von *Arabidopsis thaliana* beträgt etwa 64,2 %. Im Vergleich dazu ist das Reisgenom weniger AT-reich (Delseny, 2003), aber das der Tomate noch AT-reicher (Mizuno und Kanehisa, 1994). Der Anteil des AT-Gehalts variiert zwischen den verschiedenen Bereichen eines Genoms. AT-reiche Bereiche befinden sich häufig dort, wo die Doppelhelix getrennt werden muss. Der intergenische Bereich von *Arabidopsis thaliana* weist den höchsten AT-Gehalt auf. Es ist zu erwarten, dass der intergenische Bereich AT-reich ist, da bei der Transkription die Doppelhelix geöffnet werden muss und weniger Energie

aufgebracht werden muss, um die zwei Wasserstoffbrückenbindungen zwischen A und T zu lösen als die drei Wasserstoffbrückenbindungen zwischen G und C. Nicht kodierende Bereiche wie Introns, die 3'UTR und die 5'UTR weisen einen höheren AT-Gehalt auf als Exons. Jedoch ist der AT-Gehalt in der 5'UTR mit 61,84 % geringer als der für Introns (67,59 %) und für die 3'UTR (66,47 %). Die Analysen wurden auf die erste 5'UTR, die ersten fünf Introns, die ersten vier Exons sowie die erste 3'UTR ausgeweitet. Hierbei ist festzustellen, dass der AT-Gehalt der einzelnen Introns bzw. Exons in etwa dem der gesamten Introns bzw. Exons entspricht. Dies deutet darauf hin, dass sich dieser innerhalb der einzelnen Introns und Exons nicht wesentlich verändert. Auch bei der ersten 5'UTR sowie bei der ersten 3'UTR konnten ähnliche AT-Gehalte festgestellt werden wie für die gesamte 5'UTR bzw. 3'UTR. Da jedoch die meisten Gene UTRs aufweisen, die nicht durch Introns getrennt sind, war dies auch zu erwarten.

Nachdem der prinzipielle AT-Gehalt der einzelnen Bereiche der Gene bestimmt worden ist, wurde die Nukleotidkomposition an den einzelnen Positionen innerhalb der durchschnittlichen Länge ermittelt. Hierzu wurde ein Alignment der Sequenzen vom Beginn sowie vom Ende aus durchgeführt. Analysiert wurden die oben genannten Bereiche der 5'UTR, der Introns, Exons und der 3'UTR, jedoch wurde die Analyse auf die am höchsten nummerierten Bereiche ausgeweitet.

Die Nukleotide der ersten 5'UTR wiesen einen recht konstanten AT-Gehalt über die gesamte durchschnittliche Länge auf. Lediglich zu Beginn der ersten 5'UTR kam es zu einer Präferenz der Nukleotide AT. Weitere gemeinsame Nukleotide konnten erst bei dem Alignment der Sequenzen vom Ende der 5'UTR aufgefunden werden. Sie endet leicht bevorzugt mit AAAA, wobei das vorletzte A eine größere Bevorzugung aufweist als die übrigen drei.

Introns beginnen bevorzugt mit den Nukleotiden GTAAGT, wobei GT sehr stark konservierte Nukleotide sind. Im weiteren Verlauf sind Introns T-reich, welches sich zum Ende der Introns hin verstärkt. Die letzten Nukleotide von Introns beinhalten bevorzugt die Nukleotide T(G)(C/T)AG. Das Nukleotid T an fünftletzter Position zeigte eine größere Bevorzugung als das G an viertletzter Position. Die beiden Nukleotide C und T an drittletzter Position zeigten eine deutlichere Bevorzugung als dies für die fünftletzte Position zutrifft. Fast alle Introns enden mit AG. In der Regel beginnen Introns mit GT und enden mit AG, wobei es auch alternative Introns gibt, die AT am Anfang und AC am Ende aufweisen (Meli *et al.*, 2001). Jedoch ist der Anteil dieser alternativen Introns bei *Arabidopsis thaliana* als gering

anzusehen, da der Informationsgehalt in den erzeugten Sequenzlogos bei den ersten und letzten beiden Nukleotiden von Introns nahe bei dem maximal erreichbaren Wert liegt.

Die Analyse der Nukleotidkomposition der ersten und letzten Exons eines Gens ergab einen präferenziellen Beginn mit den Nukleotiden ATG(G), wonach es zu keinerlei Bevorzugung innerhalb der durchschnittlichen Länge mehr kam, und Exons enden bevorzugt mit T(A/G)(AG).

Das ATG stellt das bevorzugte Translations-Startcodon dar. Es konnte gezeigt werden, dass die am meisten frequentierten Nukleotide um diese Initiationssite in Monokotyledonen die Sequenz A(A/G)CC**A**$_{+1}$**TG**GC, in Eudicotyledonen A(A/C)AA**A**$_{+1}$**TG**GC (Kawaguchi und Bailey-Serres, 2005) beinhalten. In den hier vorliegenden Analysen konnte eine Bevorzugung für das Nukleotid C nach dem ATGG nicht bestätigt werden. Allerdings beziehen sich die Angaben der bevorzugten Sequenzen nicht direkt auf *Arabidopsis thaliana*. Es gibt jedoch auch nicht-ATG Startcodons, die in einem Nukleotid vom Startcodon ATG abweichen, wie CTG (Leucin) und GTG, welches in Prokaryonten auftritt (Day und Tuite, 1998). Beim *Arabidopsis* MADS-Box Gen AG konnte gezeigt werden, dass es das Startcodon ACG aufweist, welches für die Initiation der Translation *in vivo* notwendig ist (Riechmann *et al.,* 1999).

Am Ende des letzten Exons befindet sich das Stoppcodon, welches aus einem der drei Codons TAA, TAG und TGA besteht. In den hier vorliegenden Analysen konnte durch die unterschiedliche Präferenz der letzten beiden Nukleotide in Bezug auf die bevorzugt vorkommenden Stoppcodons postuliert werden, dass die beiden Codons TAA und TGA häufiger auftreten als das Codon TAG. Exons, mit Ausnahme des ersten, beginnen bevorzugt mit GT, wonach bis zum Ende keine Bevorzugung für bestimmte Nukleotide auftritt. Exons zeigen am Ende eine Bevorzugung für das Auftreten der Nukleotide A(G/A), wobei sich das letzte Exon eines Gens in der Nukleotidpräferenz unterscheidet. Innerhalb der Nukleotidsequenz von Exons war ein alternierender AT-Gehalt zu verzeichnen. Dieser bewegte sich zwischen 50 und 60 %. Dies hängt vermutlich mit der unterschiedlichen Codon Usage zusammen. Der AT-Gehalt der ersten drei Positionen beträgt 55,41 %, 49,59 % und 57,62 % (http://www.kazusa.or.jp/codon/cgi-bin/showcodon.cgi?species=3702).

Die 3'UTR zeigt keine Bevorzugung für bestimmte Nukleotide am Beginn des Bereichs. Im Verlauf der 3'UTR ist ein leichter symptotischer Anstieg des AT-Gehalts von etwa 65 % auf knapp unter 70 % zu verzeichnen. Bei einem Alignment der Sequenzen vom Ende der 3'UTR kommt es zu einen wellenförmigen Verlauf bei den letzten 30 Nukleotiden. Hierbei

überschritt der ansteigende AT-Gehalt die 70 %. Zum Ende hin erreichte der AT-Gehalt seinen Minimalwert. Am Ende war eine leichte Bevorzugung für das Nukleotid T zu verzeichnen, die während des wellenförmigen Verlaufs zunächst ab und dann wieder zunahm, wobei sich A und T gegenläufig zueinander verhielten. Im Sequenzlogo ließ sich diese leichte Präferenz für T bzw. A bestätigen, wobei der Informationsgehalt von T von der neuntletzten bis zur fünftletzten Position am größten war.

Es konnte gezeigt werden, dass sich in der 3'UTR von mRNAs AU-reiche Elemente (AREs) befinden. Das Vorhandensein dieser Elemente kann die Stabilität der mRNA dem Level der Deadenylierung, der Demaskierung oder dem 3'-5'-Abbau anpassen. Des Weiteren können diese Elemente die translationale Effizienz bestimmen und erleichtern die Kontrolle der Genexpression auf vielen Ebenen (Wilusz und Wilusz, 2004). Ein ARE hat etwa die Sequenz $(AUUUA)_n$-$(U)_n$ (Day und Tuite, 1998).

Ein Merkmal der 3'UTR ist das Auftreten eines Polyadenylierungssignals, welches bewirkt, dass an die mRNA ein Poly-A-Schwanz angehängt wird. Im tierischen System weist das Polyadenylierungssignal vorzugsweise in einem diskreten Abstand von 30-35 Nukleotiden zum Ende der 3'UTR hin die Sequenz AAUAAA auf, in selteneren Fällen auch AUUAAA. Bei pflanzlichen 3'UTRs tritt eine ähnliche Sequenz auf, die jedoch weniger gut konserviert ist und in unterschiedlichen Abständen vom Ende der 3'UTR vorkommt (Knippers, 1997). Im Rahmen der vorliegenden Analysen konnte dieses Signal nicht identifiziert werden, was vermutlich an dem nicht konservierten Abstand zum Ende der 3'UTR liegt.

4.2 Neues komplexes Modell für die Analyse von Transkriptionsfaktor-Bindungsstellen

Für die Analyse von putativen Transkriptionsfaktor-Bindungsstellen wurden pflanzliche Matrizen eingesetzt, welche genomweit nach Bindungsstellen mit der Software Patser durchsucht wurden (Hertz und Stormo, 1999). Das Ziel war, nicht nur diejenigen Bindungsstellen zu identifizieren, welche durch die in die jeweilige Matrize eingehenden Einzelsequenzen beschrieben wurden, sondern auch neue putative Bindungsstellen vorherzusagen. Bei der Erstellung der AthaMap wurde diese Methode ebenfalls eingesetzt (Steffens *et al.*, 2004, Steffens *et al.*, 2005), jedoch wurde bei den Einstellungen des Suchprogramms Patser bei der AthaMap der automatische Score verwendet, während bei den hier vorliegenden Analysen die Anzahl der putativen Bindungsstellen auf jeweils 7500 eingeschränkt worden ist. Analysen bezüglich der vorhandenen Kernsequenz bei beiden Verfahren konnten belegen, dass beide Methoden etwa ähnliche Ergebnisse lieferten, so dass davon ausgegangen werden kann, dass sich die Einschränkung der Anzahl der Bindungsstellen nicht nachteilig auf die Analysen auswirkt.

Neben dem Einsatz von Alignment Matrizen und der Software Patser gibt es noch weitere Möglichkeiten, putative Bindungsstellen vorherzusagen. Bei der online Ressource AtcisDB beispielsweise wird eine verkürzte Konsensus-Sequenz eingesetzt, wobei das Motiv dann im Promotorbereich von Genen kenntlich gemacht wird (Davuluri *et al.*, 2003). Im Rahmen der vorliegenden Arbeit wurde die Analyse nicht auf den Promotorbereich beschränkt, wodurch die Identifizierung von bislang nicht bekannten Bindungsstellen auch innerhalb von Genen möglich war. Des Weiteren kann die matrizenbasierte Suche durch die zugelassene größere Variabilität auch putative Bindungsstellen ausgeben, die in ihrer Zusammensetzung noch nicht verifiziert wurden, aber mathematisch gesehen sehr wahrscheinlich mögliche Bindungsstellen darstellen.

Um Hintergrundwerte zu erlangen wurde die Monte Carlo Methode eingesetzt. Diese Methode wird verwendet, wenn mathematische Formeln die reale Situation nicht vollständig erfassen können. Sie dient einer Approximation (Behnen und Neuhaus, 1995). Für die Erzeugung von Hintergrundwerten wurden zufällige Positionen auf dem Genom erzeugt. Die theoretische Anzahl an Bindungsstellen lässt sich nur dann mathematisch berechnen, wenn die Einflussfaktoren allesamt bekannt sind und eine diese Faktoren beschreibende Formel entwickelt wurde. Wird die Monte Carlo Methode eingesetzt, ist diese Formel nicht notwendig. Im Rahmen dieser Arbeit konnte festgestellt werden, dass bei einer zufälligen Verteilung von Bindungsstellen die Mehrzahl im downstream-Bereich von Genen auftreten und dass die Verteilung glockenkurvenförmig ist. In der Literatur ist derartiges bislang nicht beschrieben.

In den vorliegenden Analysen wurde das Vorkommen von mit pflanzlichen Matrizen detektierten putativen Bindungsstellen mit den entsprechenden Zufallsmatrizen verglichen. Hierbei stimmten die Länge der beiden Matrizen und die in die Anzahl der in die Matrize eingehenden DNA-Sequenzen überein. Der AT-Gehalt der zu erzeugenden Zufallsmatrize wurde auf den der Ursprungsmatrize eingestellt. Der Vorteil bei derartig erzeugten Vergleichsgruppen liegt darin, dass die Parameter der Ursprungsmatrize beibehalten wurden. Eine derartige Methode wurde in der Literatur bislang nicht eingesetzt. Molina und Grotewold (2005) haben DNA-Motive identifiziert und diese mit randomisierten Sequenzen verglichen. Mit der Software MEME wurden DNA-Sequenzen im Promotorbereich identifiziert und bei der Erzeugung von Vergleichswerten wurde nicht der AT-Gehalt der identifizierten Sequenzen, sondern der AT-Gehalt des intergenischen Bereichs eingesetzt. Dies führt dazu, dass die Architektur des intergenischen Bereichs, nicht aber die der identifizierten DNA-Sequenzen, in die Analysen mit eingingen.

Bei den hier vorliegenden Analysen wurden die Daten, ermittelt mit den pflanzlichen Matrizen, mit denen für nichtpflanzliche Matrizen verglichen. Des Weiteren wurden entsprechende Zufallsmatrizen eingesetzt, um den Einfluss der Parameter der

Ursprungsmatrizen auf die Verteilung zu untersuchen. Derart komplexe Analysen wurden bislang nicht veröffentlicht.

4.3 Die Verteilung von TFBSn relativ zum Translationstart ist nicht zufällig

Die Verteilung von TFBSn wurde in 50 bp Schritten relativ zum nächstgelegenen Translationsstart analysiert. Hierbei musste entschieden werden, welchem Gen eine Bindungsstelle zugeordnet werden soll, wenn diese zwischen zwei Genen lokalisiert ist. Es wurde das Kriterium des nächstgelegenen Translationsstarts gewählt, da dieser Bereich für alle Gene bekannt ist und ein allgemein gültiges Kriterium für alle Analysen benötigt wurde. Die Notwendigkeit, ein Kriterium zu verwenden, liegt in der genomweiten Analyse begründet. In der Literatur ist bislang eine derartige Zuordnung bei Verteilungen nicht durchgeführt worden. Dies hat damit zutun, dass keine genomweiten Analysen durchgeführt wurden. Berendzen *et al.* (2006) analysierten beispielsweise das Auftreten von bekannten *cis*-regulatorischen Elementen in *Drosophila melanogaster* DNA-Sequenzen in einem Bereich von -250 bp bis +50 bp relativ zum Translationsstartpunkt sowie relativ zum Transkriptionsstartpunkt. Molina und Grotewold (2005) wählten bei ihren Analysen von Motiven in *Arabidopsis thaliana* Promotoren sogar einen noch geringeren Bereich von -50 bp bis -1 bp und +1 bis +50 bp relativ zum Transkriptionsstartpunkt. In beiden Publikationen blieb eine mögliche Überlappung von Genen unberücksichtigt. Ferner wurden Hintergrundmodelle erzeugt, indem der untersuchte Bereich erweitert und überprüft wurde, wie oft das identifizierte Motiv in den erweiterten DNA-Sequenzen vorkam. Es konnte dadurch belegt werden, dass die identifizierten Motive häufiger im verkürzen Bereich auftreten. Bei genomweiten Analysen ist diese Möglichkeit des Vergleichs nicht möglich. In der vorliegenden Arbeit wurden Zufallspositionen als Hintergrundmodell verwendet.

Zunächst wurden die Zufallspositionen in dem untersuchten Bereich grafisch aufgetragen. Hierbei zeigte sich, dass es zu einer glockenförmigen Verteilung beidseitig des Translationsstartpunkts kam. Nach Analyse der 42 Datensätze konnte zwar bestätigt werden, dass die Art der Verteilungen jeweils gleich ist, jedoch konnte keine Normalverteilung nachgewiesen werden, obwohl sie visuell zu vermuten war. Dies kann an der Summe der Datensätze liegen, so dass formal keine Normalverteilung resultierte. Auffällig bei den Grafiken war, dass das Maximum im downstream-Bereich lokalisiert war. Das deutet darauf hin, dass Gene nicht gleichmäßig auf den Chromosomen verteilt sind und dass die Promotoren von zwei Genen öfters beieinander liegen, die Gene somit wahrscheinlich gegenläufig orientiert sind. Das würde dazu führen, dass eine Bindungsstelle, die sich in dem Bereich befindet, zu einem der beiden Gene zugeordnet wird, im downstream-Bereich jedoch eine größere Anzahl an Nukleotiden zum nächsten Promotor

vorliegt. Nähere Untersuchungen bezüglich des Maximums im downstream-Bereich wurden nicht durchgeführt. In der Literatur wurde bislang eine derartige Analyse, die die Verteilung von Genen auf den Chromosomen und die Orientierung der Gene miteinander in Verbindung bringt, nicht veröffentlicht.
Es erfolgte eine Einteilung der Verteilungsformen in indifferent, upstreamlastig und downstreamlastig. Als indifferent wurde ursprünglich die Verteilung der Zufallspositionen definiert, die anderen Verteilungen wurden mit ihnen verglichen. Zur Einteilung der putativen Bindungsstellen in diese Typen wurde eine Abweichung von 10 % vom indifferenten Verteilungstyp gewählt. In der Literatur, in der derartige Untersuchungen nicht veröffentlicht sind, findet sich kein Vergleichswert. Berendzen *et al.* (2006) definieren eine Signifikanz genau dann als gegeben, wenn der erhaltene Wert viermal größer als die Standardabweichung vom Hintergrundmodell ist. Jedoch ist ihre Analyse nicht vergleichbar mit der hier vorliegenden, so dass ein solches Kriterium hier keinen sinnvollen Einsatz findet. Bei den vorliegenden Analysen ist es nicht gelungen, mit Hilfe statistischer Tests gleiche Verteilungsformen von matrizenbasierenden Bindungsstellen zu identifizieren. Daher wurde ein selbst gewähltes Kriterium eingesetzt, um eine Auswertung zu ermöglichen. Die hier eingesetzten Kriterien sind recht grob gewählt, jedoch ausreichend, um einen ersten Einblick bezüglich der Verteilungsformen von matrizenbasierenden Datensätzen zu erhalten.

Bei der Auswertung wurde erwartet, dass die Bindungsstellen pflanzlicher Matrizen anders verteilt sind als die Zufallspositionen. Zudem wurden mehr putative Bindungsstellen im upstream-Bereich erwartet als im downstream-Bereich, da Bindungsstellen von Transkriptionsfaktoren vorwiegend im Promotorbereich lokalisiert sind (Reményi *et al.,* 2004). Überraschend war, dass die Bindungsstellen der pflanzlichen Matrizen vorwiegend indifferent und downstreamlastig verteilt waren. Nur etwa ein Fünftel zeigte eine upstreamlastige Verteilung. Das könnte darauf hindeuten, dass die Verteilung der Bindungsstellen einer Matrize stark vom AT-Gehalt abhängen könnte. Die Zufallsmatrizen der pflanzlichen Matrizen hingegen sollten eine Verteilung aufweisen, die den generellen Einfluss einer Matrize, aber auch den AT-Gehalt widerspiegelt. Prinzipiell sollten sie anders verteilt sein als die pflanzlichen Matrizen. Die Analyse zeigte, dass die meisten Bindungsstellen der Zufallsmatrizen für die pflanzlichen Matrizen eine downstreamlastige Verteilung aufwiesen. Da die Zufallsmatrizen verschiedene AT-Gehalte besitzen deutet dies darauf hin, dass der AT-Gehalt nicht der alleinige Parameter für die Verteilung der putativen Bindungsstellen sein kann. Beim direkten Vergleich zwischen der Verteilung von putativen Bindungsstellen der pflanzlichen Matrizen mit ihren entsprechenden Zufallsmatrizen wurde in den meisten Fällen keine Übereinstimmung in der Verteilungsform festgestellt. Dies spricht für eine Anhäufung möglicherweise biologisch relevanter Bindungsstellen bei den

pflanzlichen Matrizen, da die Zufallsmatrizen als Kontrollgruppe fungieren. Meist kam es zu einer Veränderung von der indifferenten Verteilung der Bindungsstellen pflanzlicher Matrizen hin zu einer downstreamlastigen Verteilung bei den entsprechenden Zufallsmatrizen. Das könnte darauf hindeuten, dass biologisch relevante DNA-Sequenzen mit den pflanzlichen Matrizen detektiert werden konnten. In einigen Fällen wurde eine starke Veränderung von der upstreamlastigen Verteilung hin zur downstreamlastigen Verteilung festgestellt. In der Regel traf dies auf vereinzelte Matrizen zu. Auffällig war jedoch, dass drei bZIP Matrizen diese starke Veränderung des Verteilungstyps zeigten. Das deutet darauf hin, dass die Kernsequenz, die von bZIP Faktoren erkannt wird, eine Anhäufung im Promotorbereich von Genen aufweist. Dies deckt sich mit den Analysen von Molina und Grotewold (2005), die eine Anhäufung des ACGT Motivs, welches die Kernsequenz von bZIP Faktoren darstellt, im upstream-Bereich von Genen belegt haben.

Eine weitere Kontrollgruppe waren die nichtpflanzlichen Matrizen. Hier wurde erwartet, dass die Verteilungstypen sich von pflanzlichen Matrizen unterscheiden, da davon ausgegangen wurde, dass mit den nichtpflanzlichen Matrizen keine funktionalen Bindungsstellen detektiert werden können. In der Tat wies die überwiegende Anzahl eine downstreamlastige Verteilung auf und damit lag ein Unterschied zu den pflanzlichen Matrizen vor. Ein großer Unterschied war in der upstreamlastigen Verteilung zu erkennen. Nur bei einer Matrize konnte dieser Verteilungstyp identifiziert werden. Prinzipiell deutet das darauf hin, dass mit dieser Methode möglicherweise neue Transkriptionsfaktorklassen, die in Pflanzen vorkommen, identifiziert werden könnten. Berendzen *et al.* (2006) belegten, dass in Promotorsequenzen von *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster* und *Saccharomyces cerevisiae* gemeinsame Motive auftreten, was für einen generellen Aufbau von bestimmten Promotoren spricht. Die in der vorliegenden Arbeit auffällige Matrize gehört zur Klasse Fork head und zeichnet sich durch einen sehr hohen AT-Gehalt aus. Die Klasse Fork head kommt laut Riechmann *et al.* (2000) nicht in Pflanzen vor.
Von den Zufallsmatrizen für die nichtpflanzlichen Matrizen wurde erwartet, dass sie sich ähnlich wie ihre entsprechenden nichtpflanzlichen Matrizen verhalten. Aber dies war nicht immer der Fall. Es zeigte sich bei den Zufallsmatrizen eine noch deutlichere Bevorzugung für die Lokalisierung der Bindungsstellen im downstream-Bereich. Die entsprechende Zufallsmatrize des nichtpflanzlichen Fork head Transkriptionsfaktors wies jedoch ebenfalls eine Bevorzugung für den upstream-Bereich auf, so dass dieser Effekt vermutlich auf dem AT-Gehalt der Matrizen beruhen könnte. Beim direkten Vergleich zwischen den nichtpflanzlichen Matrizen mit den entsprechenden Zufallsmatrizen wurden auch Veränderungen in der Verteilungsform festgestellt. In der Regel war dies von der indifferenten Verteilung hin zu einer downstreamlastigen Verteilung. Dies traf beispielsweise

auf die Bindungsstellen der Matrizen von COUP, PPAR- alpha, HAP3, E2F, Pax-3, Pax-6, Dl und Brachyury zu. Diese Ergebnisse deuten darauf hin, dass möglicherweise neue Transkriptionsfaktor-Klassen, die bislang noch nicht in Pflanzen identifiziert worden sind, vorhergesagt werden könnten.

Zusammengefasst lässt sich feststellen, dass Zufallspositionen eine indifferente Verteilung aufweisen, aber dass Zufallsmatrizen für pflanzliche sowie für nichtpflanzliche Matrizen eine starke Bevorzugung für eine downstreamlastige Verteilung zeigen. Dies hängt nicht allein vom AT-Gehalt ab, sondern scheint mit im Matrizencharakter begründet zu sein. Auch wenn bei den Zufallsmatrizen keine realen Bindungsstellen zugrunde liegen, werden doch bestimmte DNA-Sequenzen identifiziert, welche dann jedoch bevorzugt im downstream-Bereich vorkommen. Das könnte damit zusammenhängen, dass der Selektionsdruck auf bestimmte regulatorische Sequenzen dort nicht vorherrscht und eine größere Variabilität von DNA-Sequenzen vorliegt. Diesbezügliche Untersuchungen in der Literatur gibt es bislang noch nicht.

4.4 Bindungsstellen für bestimmte Transkriptionsfaktoren zeigen spezifische Anreicherungen in einzelnen Genbereichen

Die Lokalisierung der Zufallspositionen und der matrizenbasierenden putativen TFBSn wurden im intergenischen Bereich, in der 5'UTR, den Introns, den Exons und der 3'UTR erfasst.

Der Vergleich der prozentualen Anzahl der Zufallspositionen in den einzelnen Bereichen innerhalb von Genen und im intergenischen Bereich deckt sich mit der prozentualen Anzahl der Nukleotide, die in diesen Bereichen lokalisiert sind. Das deutet darauf hin, dass die Zufallspositionen tatsächlich zufällig und damit gleichmäßig auf den Chromosomen verteilt sind und ein geeignetes Hintergrundmodell liefern. Die Werte für die Zufallspositionen wurden mit denen der matrizenbasierenden putativen Bindungsstellen verglichen.

Im intergenischen Bereich wird erwartet, dass TFBSn dort vorwiegend lokalisiert sind (Reményi *et al.,* 2004). Überraschend ist, dass die Anzahl der pflanzlichen putativen TFBSn niedriger war als die der Zufallspositionen. Dies deutet auf eine mögliche Gegenselektion im intergenischen Bereich hin. Die Anzahl der Bindungsstellen der Zufallsmatrizen sowie die der nichtpflanzlichen Matrizen war noch geringer als dies bei den pflanzlichen Matrizen der Fall war. Das könnte darauf hindeuten, dass sich zwei Effekte überlagern, einerseits eine

Gegenselektion in Bezug auf zufällige DNA-Sequenzen, andererseits das Detektieren von biologisch relevanten TFBSn bei den pflanzlichen Matrizen.

Generell treten biologisch relevante TFBSn nicht nur im intergenischen Bereich, sondern z.B. auch in der 5'UTR sowie der 3'UTR auf (Feldbrügge *et al.,* 1994, Hyder *et al.,* 2000). Die angegebenen Literaturstellen beziehen sich nicht auf *Arabidopsis*, sondern auf die Petersilie bzw. die Ratte. Bezüglich der 5'UTR kamen Zufallspositionen und matrizenbasierende pflanzliche TFBSn etwa gleich häufig vor. Auch die Anzahl der Bindungsstellen der Zufallsmatrizen und der nichtpflanzlichen Matrizen unterschied sich nicht von den Zufallspositionen. Das könnte darin begründet liegen, dass generell wenige Nukleotide in der 5'UTR lokalisiert sind. In der 3'UTR war die Anzahl der Zufallspositionen im Vergleich zu den matrizenbasierenden TFBSn leicht erhöht. Jedoch sind die Werte generell recht niedrig, was auch in der geringen Anzahl der Nukleotide in diesem Bereich begründet liegen dürfte.

Die Lokalisierung von TFBSn in Introns ist mehrfach beschrieben worden (Sieburth und Meyerowith, 1997, Tang und Perry, 2003). Daher sind putativ funktionale TFBSn in Introns zu erwarten. Im Vergleich zu den Zufallspositionen traten weniger pflanzliche Bindungsstellen auf, was auf eine Gegenselektion von zufälligen DNA-Sequenzen hindeutet.

Das mögliche Auftreten von TFBSn in einem Exon ist von Merriam und Chess (2007) beschrieben worden. Die prozentuale Anzahl der Zufallspositionen ist im Vergleich zu denen der pflanzlichen Matrizen geringer. Das deutet auf keine Gegenselektion von Bindungsstellen in diesem Bereich hin. Dies könnte daran liegen, dass die Variabilität der Nukleotidzusammensetzung in Exons größer ist als in anderen Genbereichen. Im ersten Exon war die Anzahl der Zufallspositionen sowie die der pflanzlichen Matrizen größer als in den darauf folgenden drei Exons. Das hängt sehr wahrscheinlich damit zusammen, dass viele Gene lediglich ein Exon besitzen und dass die Anzahl der Nukleotide, die im ersten Exon liegen, deutlich größer ist als in den darauf folgenden.

Im weiteren Verlauf der Analysen wurden von den Bindungsstellen jeder pflanzlichen Matrize in den untersuchten Bereichen Abweichungsfaktoren von den Zufallspositionen gebildet und grafisch aufgetragen. Dies diente dem direkten Vergleich, um mögliche biologisch relevante Anhäufungen von Bindungsstellen in einem bestimmten Bereich feststellen zu können.

Im intergenischen Bereich werden putative TFBSn von pflanzlichen Matrizen erwartet (Reményi *et al.,* 2004). Der Vergleich mit den Zufallspositionen konnte zeigen, dass es in vielen Fällen zu positiven Abweichungsfaktoren kam. Es wurden also mehr putative TFBSn

ermittelt als anhand der Zufallspositionen erwartet wurden. Alle Matrizen, deren Transkriptionsfaktoren zur Klasse MADS gehören, wiesen Anhäufungen auf. In der Literatur gibt es einige Beispiele für die Lokalisierung von MADS TFBSn im Promotorbereich (Tang und Perry, 2003, Busch *et al.*, 1999). Beide Matrizen der Klasse bHLH zeigten mehr Bindungsstellen im intergenischen Bereich als die Zufallspositionen. Dies deckt sich mit den Literaturangaben, dass der Transkriptionsfaktor PIF3 an Promotorsequenzen seiner Zielgene bindet (Martinez-Garcia *et al.,* 2000). Gleiches gilt für die Klasse Trihelix, C2H2(Zn) und TBP, wobei jeweils nur eine Matrize für die entsprechende Transkriptionsfaktor-Klasse eingesetzt wurde. GT1, der Trihelix Transkriptionsfaktor, bindet an eine DNA-Sequenz im Promotor des Gens Tdc (Ouwerkerk *et al.*, 1999). Im Fall von ID1, einem Transkriptionsfaktor der Klasse C2H2(Zn), ist zwar bekannt, dass er den Transkriptionsfaktor DLF1 reguliert. Wo die Bindungsstelle lokalisiert ist, ist jedoch derzeit nicht bekannt (Muszynski *et al.*, 2006). Bei TBP ist bekannt, dass der Faktor im Promotorbereich von Genen bindet (Shahmuradov *et al.,* 2003). Jedoch wurden für die Matrize, die bei den hier vorliegenden Analysen eingesetzt wurde, nur *in silico* Sequenzen pflanzlicher TATA-Boxen aus Promotoren eingesetzt (Shahmuradov *et al.,* 2003). Die Promotoren vieler Gene weisen eine TATA-Box auf, von daher kann davon ausgegangen werden, dass der positive Abweichungsfaktor von den Zufallspositionen auf eine erwartete Anreicherung von Bindungsstellen im Promotorbereich von Genen zurückzuführen ist. Bei bZIP Transkriptionsfaktoren ist einerseits bekannt, dass sie DNA-Sequenzen im Promotorbereich erkennen (Singh, 1998), andererseits wurde auch schon auf bioinformatischem Weg gezeigt, dass im Promotorbereich von Genen gehäuft bZIP Motive vorkommen (Molina und Grotewold, 2005). Bei den in der vorliegenden Arbeit ermittelten Ergebnissen zeigten nur drei der acht eingesetzten bZIP Matrizen gehäuft Bindungsstellen im Promotorbereich. Bei den Analysen von Molina und Grotewold (2005) wurden Motive mit einer Länge von 10 Nukleotiden identifiziert. Des Weiteren bezog sich deren Analyse auf 12749 Gene. In den hier vorliegenden Analysen waren die eingesetzten Matrizen länger und die Anzahl der Gene betrug 29993, was zu einer detaillierten Analyse sowie einer stärkeren Selektion von Bindungsstellen führte. Von vier eingesetzten Matrizen für die Klasse HD-ZIP zeigten drei positive Abweichungsfaktoren von den Zufallspositionen. Hier ist von einer bevorzugten Anreicherung von DNA-Sequenzen, an die HD-ZIP Faktoren binden können, im intergenischen Bereich auszugehen. Von Homeodomain (HD) Transkriptionsfaktoren ist bekannt, dass diese an Promotorsequenzen ihrer Zielgene binden (Chan *et al.*, 1998). Bei den beiden Matrizen für Faktoren der Klasse TCP zeigte nur eine einen positiven Abweichungsfaktor. Kosugi und Ohashi (2002) konnten belegen, dass TCP Transkriptionsfaktoren an Promotorsequenzen ihrer Zielgene binden können. Im Fall von Transkriptionsfaktoren der Klasse MYB, für die insgesamt acht Matrizen eingesetzt wurden,

konnte lediglich bei zwei Matrizen eine Häufung von Bindungsstellen im intergenischen Bereich festgestellt werden. Bekannt ist, dass MYB Transkriptionsfaktoren an DNA-Sequenzen in Promotorbereich ihrer Zielgene binden (Solano *et al.*, 1995, Gubler *et al.*, 1999). Im intergenischen Bereich zeigten jedoch nicht alle pflanzliche Matrizen positive Abweichungsfaktoren von den Zufallspositionen. Das kann daran liegen, dass bestimmte regulatorische DNA-Sequenzen nicht gehäuft auftreten und somit auch nicht mit dieser Methode detektiert werden können.

Für die Bindungsstellen der Zufallsmatrizen der pflanzlichen Matrizen wurden ebenfalls Abweichungsfaktoren von den Zufallspositionen gebildet. Insgesamt wurden weniger häufig positive Abweichungsfaktoren identifiziert. Dies ist ein Indiz dafür, dass es sich bei den pflanzlichen putativen TFBSn gehäuft um biologisch relevante DNA-Sequenzen handeln könnte. Der direkter Vergleich der Abweichungsrichtung (positiv-negativ) der pflanzlichen Matrize mit ihrer entsprechenden Zufallsmatrize ergab, dass bei einigen pflanzlichen Matrizen eine Änderung der Abweichungsrichtung im Vergleich zu ihrer entsprechenden Zufallsmatrize zu verzeichnen war. Das deutet darauf hin, dass der AT-Gehalt grundsätzlich mit in die Verteilung von Transkriptionsfaktoren eingehen kann, jedoch nicht der alleinige Grund für positive Abweichungsfaktoren von den Zufallspositionen darstellt.

Bei den nichtpflanzlichen Matrizen kam es ebenfalls zu positiven Abweichungsfaktoren von den Zufallspositionen, jedoch traten diese weitaus weniger häufig auf. Das deutet darauf hin, dass das Vorkommen positiver Abweichungsfaktoren von den nichtpflanzlichen Matrizen im intergenischen Bereich vorwiegend auf dem AT-Gehalt beruhen könnte. Der direkte Vergleich zwischen den Ergebnissen für die nichtpflanzlichen Matrizen und ihren entsprechenden Zufallsmatrizen ergab nur in einem Fall keine übereinstimmende positive Abweichung von den Zufallspositionen (Transkriptionsfaktor E2F). Der Transkriptionsfaktor E2F gehört laut TRANSFAC© zur Klasse Fork head (Matys *et al.*, 2003) und bislang ist nicht bekannt, dass Transkriptionsfaktoren dieser Klasse in Pflanzen vorkommen, es könnte sich jedoch um einen Hinweis auf ein neuartiges Vorkommen von Transkriptionsfaktor-Klassen in Pflanzen handeln. Es könnte jedoch sein, dass die Zuordnung des Faktors zu seiner Klasse überdacht werden könnte, da der Faktor eine E2F/DP Domäne aufweist, was ein Merkmal der Klasse E2F/DP ist, welche in Pflanzen vorkommt (Stevens *et al.*, 2002).

Die positiven Abweichungsfaktoren der matrizenbasierenden Daten in der ersten 5'UTR zeigten kein einheitliches Bild. Prinzipiell ist in der Literatur das Auftreten von TFBSn in der 5'UTR von Genen bekannt (Feldbrügge *et al.*, 1994, Apfel *et al.*, 1994, Farnham und Means, 1990). Bei den hier eingesetzten Matrizen ist jedoch eine derartige Lokalisierung noch nicht

beschrieben worden. Bei einigen pflanzlichen Matrizen konnten putative TFBSn identifiziert werden, die häufiger vorkommen als die Zufallspositionen. Generell lässt sich feststellen, dass positive Abweichungsfaktoren von den Zufallspositionen bei den Zufallsmatrizen häufiger auftreten als bei den pflanzlichen und nichtpflanzlichen Matrizen. Die Ergebnisse deuten darauf hin, dass unter Einsatz von Matrizen bestimmte DNA-Sequenzen, die einen bestimmten AT-Gehalt aufweisen, häufiger detektiert werden als spezielle DNA-Sequenzen, die eine vorgegebene Kernsequenz aufweisen.

Bei der Analyse von Abweichungsfaktoren von den Zufallspositionen in den ersten fünf Introns wiesen die pflanzlichen Matrizen häufig gleiche Abweichungsrichtungen auf, die in den untersuchten Introns beibehalten wurden. Das könnte darauf hindeuten, dass bestimmte Introns teilweise eine ähnliche DNA-Sequenz aufweisen.

In der Literatur ist das Vorkommen von TFBSn in Introns in zahlreichen Fällen beschrieben worden (Jeong et al., 2006, Sieburth und Meyerowitz, 1997, Busch et al., 1999, Tang und Perry, 2003). Bei den für die vorliegende Arbeit eingesetzten Matrizen ist für den Faktor AG bekannt, dass AP2, LUC und ANT im Intron des Gens AG DNA-Sequenzen erkennen, an sie binden und die Transkription regulieren (Sieburth und Meyerowitz, 1997). Für die restlichen in der vorliegenden Arbeit eingesetzten Transkriptionsfaktoren ist derartiges noch nicht beschrieben worden. Bei der Matrize des Faktors ANT ist in allen fünf Introns eine negative Abweichung von den Zufallspositionen zu verzeichnen. Das kann darauf zurückzuführen sein, dass regulatorische Sequenzen nicht gehäuft in Introns vorkommen. In einigen Fällen kam es bei den pflanzlichen Matrizen im Vergleich zu ihren entsprechenden Zufallsmatrizen zu Änderungen in der Abweichungsrichtung. Das deutet auf einen biologischen Einfluss bei den durch die pflanzlichen Matrizen detektierten putativen TFBSn hin.

Im Vergleich zu den putativen TFBSn der pflanzlichen Matrizen zeigten die der Zufallsmatrizen der pflanzlichen Matrizen weniger häufig positive Abweichungsfaktoren von den Zufallspositionen. Ähnliches konnte für die nichtpflanzlichen Matrizen und deren Zufallsmatrizen festgestellt werden. Die Zufallsmatrizen der nichtpflanzlichen Matrizen wiesen weniger häufig positive Abweichungsfaktoren auf als die nichtpflanzlichen Matrizen. Auch die Abweichungsrichtung änderte sich selten in den untersuchten Introns. Dies deutet darauf hin, dass die Nukleotidzusammensetzung von bestimmten Introns in bestimmten Bereichen ähnlich sein könnte. Diesbezügliche Untersuchungen wurden bislang in der Literatur nicht veröffentlicht.

In den untersuchten Exons traten die meisten positiven Abweichungsfaktoren von den Zufallspositionen auf. Das traf auf alle matrizenbasierenden TFBSn zu. Dies deutet nicht auf eine Gegenselektion bezüglich bestimmter DNA-Sequenzen hin. Im Verlauf der Exons zeigten die TFBSn der pflanzlichen und nichtpflanzlichen Matrizen weniger häufig Änderungen in ihrer Abweichungsrichtung als die Zufallsmatrizen. Das deutet darauf hin, dass konservierte DNA-Sequenzen, die durch die pflanzlichen bzw. nichtpflanzlichen Matrizen beschrieben wurden, einen einheitlicheren Effekt auf das verstärkte oder verminderte Vorhandensein der putativen Bindungsstellen in den Exons ausüben als zufällige DNA-Sequenzen. In der Literatur ist in einem Fall beschrieben worden, dass Bindungsstellen im Exon vorkommen können (Merriam und Chess, 2007).

In der ersten 3'UTR kam es bei den pflanzlichen Matrizen in wenigen Fällen zu positiven Abweichungsfaktoren von den Zufallspositionen. Ein ähnliches Ergebnis konnte bei den Zufallsmatrizen der pflanzlichen Matrizen verzeichnet werden. In seltenen Fällen kam es zu einer Änderung der Abweichungsrichtung der TFBSn der pflanzlichen Matrizen im Vergleich zu ihren entsprechenden Zufallsmatrizen. Das deutet darauf hin, dass es eine mögliche Gegenselektion sowohl von bestimmten, durch die pflanzlichen Matrizen beschriebenen, DNA-Sequenzen als auch bezüglich des AT-Gehalts kam. Ähnliche Ergebnisse wurden für die nichtpflanzlichen Matrizen und deren dazugehörigen Zufallsmatrizen ermittelt, was die mögliche Ursache einer Gegenselektion bekräftigt. In der Literatur ist bekannt, dass in der 3'UTR von Genen regulatorische DNA-Sequenzen liegen, an die Transkriptionsfaktoren binden (Banks *et al.*, 2000, Hyder *et al.*, 2000). Von den Transkriptionsfaktoren, die in der vorliegenden Arbeit eingesetzt wurden, ist derartiges bislang nicht bekannt.

Aufgrund der obigen Auswertung wurden Auffälligkeiten protokolliert. Als auffällig wurden Fälle angesehen, in denen pflanzliche Matrizen eine im Vergleich zu den restlichen drei matrizenbasierenden Ergebnissen (Zufallsmatrizen für die pflanzlichen Matrizen, nichtpflanzliche Matrizen und die Zufallsmatrizen der nichtpflanzlichen Matrizen) unterschiedliche Abweichungsrichtung zeigten. Mit den resultierenden Kandidaten bieten sich weitergehende Analysen in Bezug auf die biologische Relevanz der putativen TFBSn in den entsprechenden Bereichen innerhalb von Genen und im intergenischen Bereich an. Vergleichbare Analysen gibt es in der Literatur bislang noch nicht.

4.5 Eingrenzung der potentiellen Promotorbereiche durch Abweichungsfaktoren von den Zufallspositionen

Es wurden Verteilungsanalysen mit den Zufallspositionen sowie den putativen TFBSn der pflanzlichen Matrizen, der Zufallsmatrizen der pflanzlichen Matrizen, der nichtpflanzlichen Matrizen sowie der Zufallsmatrizen der nichtpflanzlichen Matrizen im intergenischen Bereich durchgeführt. Diese Analysen wurden in 1 bp Schritten relativ zum nächstgelegenen Gen durchgeführt.

Bei den Zufallspositionen konnte eine Anreicherung von Zufallspositionen, die im upstream-Bereich von Genen lokalisiert sind, festgestellt werden, welche mit abnehmender Distanz zum Beginn des Gens zunahm. Im downstream-Bereich, das heißt, dass sich die Zufallspositionen downstream vom Ende des Gens befanden, traten weniger Zufallspositionen auf. Dass es zu einer prinzipiellen Anreicherung von Positionen im upstream-Bereich von Genen kommt, kann damit zusammenhängen, dass das Genom dicht gepackt ist (Haas *et al.*, 2005), und dass Bindungsstellen, die im intergenischen Bereich im downstream-Bereich eines Gens, aber im upstream-Bereich des nächstgelegenen Gens lokalisiert sind, zu diesem gezählt werden. Die Zufallspositionen lieferten Hintergrunddaten unter Voraussetzung einer gleichmäßigen Verteilung der Zufallspositionen im Genom. Mit der hier vorliegenden Arbeit vergleichbare Analysen gibt es in der Literatur bislang nicht.

Im weiteren Verlauf der Analysen erfolgte die Bildung von Abweichungsfaktoren der Bindungsstellen der pflanzlichen Matrizen, der Zufallsmatrizen der pflanzlichen Matrizen, der nichtpflanzlichen Matrizen sowie der Zufallsmatrizen der nichtpflanzlichen Matrizen von den Zufallspositionen.

Überraschend war, dass es im upstream-Bereich von Genen zu negativen Abweichungsfaktoren von den Zufallspositionen kam, welche mit abnehmender Distanz zum Gen tendenziell größer wurden. Dieses Resultat konnte für alle Matrizen festgestellt werden, nicht nur für die pflanzlichen Matrizen. Das deutet darauf hin, dass es eine Gegenselektion für bestimmte DNA-Sequenzen in diesem Bereich geben könnte. Im upstream-Bereich von Genen sind die Promotoren lokalisiert (Strasburger, 2002). Die mögliche Gegenselektion könnte mit der Promotorlänge korrelieren. Aus diesem Grund wurden die Bereiche mit negativen Abweichungsfaktoren von den Zufallspositionen ermittelt. Hierbei wurden die Werte für Transkriptionsfaktoren, deren dazugehörige Gene einen annotierten Transkriptionsstartpunkt aufweisen und Transkriptionsfaktoren, für deren dazugehörige Gene lediglich der Translationsstartpunkt annotiert ist, getrennt voneinander betrachtet.

Bei TFBSn, für deren dazugehörige Gene ein Transkriptionsstartpunkt annotiert ist, lagen die durchschnittlichen Distanzen mit negativen Abweichungsfaktoren zwischen -413,11 bp bis -422,19 bp relativ zum Transkriptionsstart und reichten bis -3,81 bp bis -2,2 bp an ihn heran. Überraschend ist, dass die Werte, unabhängig davon, ob pflanzliche Matrizen eingesetzt wurden oder Zufallsmatrizen bzw. nichtpflanzliche Matrizen, einheitlich sind. Das deutet darauf hin, dass möglicherweise genau in dem Bereich von durchschnittlich -405 bp bis -3 bp eine Gegenselektion von nicht regulatorischen DNA-Sequenzen die Ursache für diese negativen Abweichungsfaktoren sein könnte. Ein eukaryontischer Kernpromotor ist etwa 70 Nukleotide lang und reicht bis zum Transkriptionsstartpunkt (Molina und Grotewold, 2005). Als Kernpromotor werden Bindungsstellen für basale Transkriptionsfaktoren angesehen. Laut Strasburger (2002) umfasst die Länge des Promotors die ersten 200 bp upstream vom Transkriptionsstartpunkt. In den hier vorliegenden Analysen wurden Enhancer Elemente ebenfalls zum Promotor gezählt. Enhancer Elemente können upstream von Genen, in der Nähe des Kernpromotors, aber auch einige kb downstream vom Gen lokalisiert sein (Pedersen *et al.*, 1999). Bei *Arabidopsis thaliana* sind die Promotorlängen incl. Enhancer Elemente keinesfalls einheitlich. Bei Athena, einer Datenbank für *Arabidopsis thaliana* Promotoren, beträgt die Standardeinstellung z.B. 1000 bp upstream vom Transkriptionsstartpunkt und lässt sich auf Werte zwischen 250 bp und 3000 bp variieren (O'Connor *et al.*, 2005). Werden die Einstellungen von O'Connor *et al.* (2005) mit den Ergebnissen in der hier vorliegenden Arbeit verglichen, deutet es darauf hin, dass es sich bei den identifizierten Bereichen um Promotorbereiche handeln könnte.

Bei den Analysen mit TFBSn, deren dazugehörige Gene lediglich einen Translationsstart aufweisen, stimmten die Distanzen mit negativen Abweichungsfaktoren von den Zufallspositionen in etwa überein. Es traten Distanzen von -338,63 bp bis -377,95 bp und -11,44 bp bis -12,81 bp auf. Auch hier war es unerheblich, ob pflanzliche Matrizen oder Zufallsmatrizen bzw. nichtpflanzliche Matrizen eingesetzt worden waren. Die Distanzen waren jedoch geringer als dies beim Vorhandensein des Transkriptionsstartpunkts der Fall war. Prinzipiell deutet auch dieses Resultat auf eine Gegenselektion von bestimmten nicht regulatorischen DNA-Sequenzen im Promotorbereich der entsprechenden Gene hin. Der Unterschied zwischen den Resultaten bei einem vorhandenen Transkriptionsstartpunkt und dem lediglich annotierten Translationstartpunkt könnte in einem durchschnittlich kürzeren Promotor begründet sein. Es könnte aber auch sein, dass der Bezugspunkt des Translationsstarts ungenauer ist, weil bei den entsprechenden Genen keine UTR annotiert ist, die in den meisten Fällen noch erwartet wird. Es kann davon ausgegangen werden, dass viele Gene, die nur einen Translationsstartpunkt aufwiesen, Transposons oder Pseudogene

usw. sein könnten, so dass die Promotoren möglicherweise anders aufgebaut sein könnten als dies für normale proteinkodierende Gene der Fall ist.

Die Grafiken mit den Abweichungsfaktoren der putativen Bindungsstellen von den Zufallsmatrizen wurden näher untersucht. Hierbei wurde ein direkter Vergleich zwischen den putativen TFBSn der pflanzlichen bzw. nichtpflanzlichen Matrize und der entsprechenden Zufallsmatrize durchgeführt, indem Überlagerungsgrafiken erzeugt wurden.

Bei den pflanzlichen Matrizen kam es in einigen Fällen zu Distanzbereichen, innerhalb derer die Anzahl der putativen pflanzlichen TFBSn deutlich über dem Niveau der entsprechenden Zufallsmatrize lag. Bei den nichtpflanzlichen Matrizen konnten derartige Distanzbereiche nicht festgestellt werden. Das deutet darauf hin, dass es in bestimmten Promotoren zu einer prinzipiellen Anhäufung dieser DNA-Sequenzen kommt. Für diese Vermutung spricht, dass z.B. die beiden eingesetzten bHLH Matrizen diese Auffälligkeitsbereiche aufwiesen. bHLH Faktoren erkennen vorwiegend die Kernsequenz ACGT (Martinez-Garcia *et al.*, 2000). Molina und Grotewold (2005) haben belegt, dass es in Promotoren eine Anhäufung der Sequenz ACGT gibt. bZIP Matrizen erkennen in der Regel ebenfalls die Kernsequenz ACGT (Jakoby *et al.*, 2002). Von den in den vorliegenden Analysen eingesetzten bZIP Matrizen traten in den meisten Fällen derartige Anhäufungen von putativen Bindungsstellen auf. Bei den entsprechenden Zufallsmatrizen fehlten diese. Daher deutet die Anreicherung auf eine mögliche biologische Relevanz hin. Im Falle der Matrize für TBP, bei deren Bindungsstellen eine Anhäufung im Promotorbereich erwartet wird (Knippers, 1997), konnte im Rahmen dieser Untersuchungen ein Bereich festgestellt werden, innerhalb dessen es zu mehr Bindungsstellen kam als bei der entsprechenden Zufallsmatrize. Des Weiteren waren bei beiden Matrizen, deren dazugehörige Transkriptionsfaktoren der Klasse TCP angehören, ebenfalls Bereiche mit einer Anhäufung von putativen TFBSn aufzufinden, wobei Kosugi und Ohashi (2002) belegten, dass Transkriptionsfaktoren dieser Klasse an Promotorsequenzen binden.
Zusammengefasst deuten die Ergebnisse darauf hin, dass mit den entsprechenden Matrizen weiterführende Analysen durchgeführt werden können, um festzustellen, ob es sich bei den Anhäufungen tatsächlich um die entsprechenden Gene regulierende TFBSn handelt. Dafür spricht ebenfalls, dass derartige Anreicherungen nur bei den pflanzlichen Matrizen auftraten, nicht jedoch bei den restlichen Matrizen. Da die zu den Bereichen gehörenden TFBSn jedoch so zahlreich auftreten, könnten beispielsweise Stichproben verwendet werden. Im Rahmen dieser Arbeit wurden keine Untersuchungen mit molekularbiologischen Methoden durchgeführt.

Bei den pflanzlichen und nichtpflanzlichen Matrizen kam es zu einzelnen Distanzen, die eine vermehrte Lokalisierung von putativen TFBSn aufwiesen. Dies kam bei bZIP Matrizen, MYB Matrizen, einer HD-ZIP Matrize, der Matrize des Faktors CBF sowie HD-Knotted vor. Generell ist bei allen Faktoren dieser Klassen die Bindung an DNA-Sequenzen im Promotor bekannt (Jakoby *et al.*, 2002, Solano *et al.*, 1995, Gubler *et al.*, 1999, Chan *et al.*, 1998, Strasburger, 2002). Die einzelnen Distanzen mit diesen Auffälligkeiten traten auch bei nichtpflanzlichen Matrizen auf. Daher ist ungewiss, ob diese Anhäufungen auf dem Zufall beruhen oder ob auch die entsprechenden Transkriptionsfaktoren, deren Klassen bislang in Pflanzen nicht identifiziert worden sind, möglicherweise neue pflanzliche Klassen darstellen könnten. Die Klassen der Transkriptionsfaktoren, mit deren Matrizen diese Auffälligkeit aufgefunden worden sind, sind CH-homeo, T-Box, CC (rec), Fork head, GCM, histone fold, C6 und paired.

Wurden bei den Analysen nur diejenigen putativen TFBSn einbezogen, deren dazugehörigen Gene nur einen annotierten Translationsstartpunkt aufwiesen, wurde bei den pflanzlichen Matrizen nur beim Faktor TBP ein Distanzbereich festgestellt, innerhalb dessen es zu einer Anhäufung von Bindungsstellen kam, die bei der dazugehörigen Zufallsmatrize fehlte. Im Fall des TBP wird eine Anhäufung erwartet (Shahmuradov *et al.,* 2003). Bei den nichtpflanzlichen Matrizen wurde kein Distanzbereich mit einer Anhäufung von putativen Bindungsstellen aufgefunden. Bei den pflanzlichen Matrizen zeigten einige Matrizen einzelne Distanzen, bei denen die Anzahl der putativen Bindungsstellen größer war als dies bei den entsprechenden Zufallsmatrizen der Fall war. Diese Auffälligkeiten wurden bei den putativen Bindungsstellen der Faktoren TBP, CBF und WRKY aufgefunden sowie bei mehreren Matrizen der Klassen bZIP, HD-ZIP und MYB und einer Matrize der Klasse TCP. In der Literatur wurden Analysen vorwiegend in Bezug zum Transkriptionsstartpunkts durchgeführt, weil sich dadurch genauere Analysen bezüglich des Promotorbereichs durchführen lassen (Molina und Grotewold, 2005, Berendzen *et al.*, 2006). Generell stellen diese Matrizen jedoch potenzielle Kandidaten für weitergehende Analysen dar. Bei einigen nichtpflanzlichen Matrizen kamen einzelne Distanzen vor, bei denen die Anzahl der putativen Bindungsstellen die der Zufallsmatrizen überragte. Auch hier ist es möglich, dass dieses Vorkommen ein Resultat des Zufalls sein könnte oder dass die entsprechenden Klassen TEA, histone fold, paired und paired-homeo möglicherweise neue pflanzliche Transkriptionsfaktor-Klassen darstellen könnten.

4.6 Gemeinsame Expression von putativen Zielgenen und den entsprechenden Transkriptionsfaktoren ermittelt durch P-Werte und Genevestigator

Bei den bisherigen Analysen wurden Vergleiche der TFBSn pflanzlicher bzw. nichtpflanzlicher Matrizen mit denen der entsprechenden Zufallsmatrizen und mit den Zufallspositionen durchgeführt. Die Resultate lieferten Hinweise bezüglich Anreicherungen oder auch Abreicherungen in bestimmten Bereichen innerhalb von Genen und im intergenischen Bereich. Im weiteren Verlauf wurde der P-Wert als statistisches Signifikanzkriterium eingesetzt, um festzustellen, ob die Anzahl pflanzlicher TFBSn in einzelnen Bereichen innerhalb von Genen und im intergenischen Bereich auf Grund des Zufalls zu erwarten ist. In der Literatur werden Wahrscheinlichkeitswerte verwendet, um Hinweise bezüglich der Zufälligkeit des Ergebnisses zu erhalten. Hannenhalli und Levy (2002) verglichen die Anzahl an Kolokalisationen von TFBSn mit der durch den Zufall resultierenden Anzahl durch Verwendung der Monte Carlo Methode. Hierbei wurden Zufallspositionen generiert, welche dann statistisch mit den tatsächlichen Kolokalisationen verrechnet wurden. Hertz und Stormo (1999) berücksichtigten bei der Berechnung des Informationsgehalts einer Matrize die erwartete Anzahl an Nukleotiden anhand des Vorkommens jedes Nukleotids im Genom. Beim BLAST (Basic Local Alignment Search Tool) wird angegeben, wie wahrscheinlich es ist, dasselbe Ergebnis mit zufällig generierten Sequenzen zu erhalten (Altschul *et al*., 1990). Hierbei werden die resultierenden P-Werte (E-Werte) der identifizierten DNA-Sequenzen lediglich nach der Größe sortiert und angegeben, eine tatsächliche Bewertung der Signifikanzen erfolgt nicht (Altschul *et al*., 1997). In der vorliegenden Arbeit wurden P-Werte nach Bartsch (1994) berechnet. Anschließend wurden die 10 signifikantesten Anhäufungen von TFBSn an bestimmten Distanzen in den einzelnen Bereichen innerhalb der Gene und im intergenischen Bereich identifiziert. Die entsprechenden putativen Gene der Transkriptionsfaktoren wurden in Genevestigator eingegeben (Zimmermann *et al*., 2005). Weil das Genom von *Arabidopsis thaliana* durchsequenziert und archiviert ist (The *Arabidopsis* Genome Initiative, 2000), ist es möglich, die Genexpression genomweit unter Einsatz von Microarrays zu untersuchen (Zimmermann *et al*., 2004). Microarray-Daten werden vielfach eingesetzt und dienen dazu, z.B. eine spezifische Genexpression unter bestimmten Wachstumsbedingungen oder unter Stress zu beobachten (Grennan, 2006, Seki *et al*., 2002). In der vorliegenden Arbeit wurden putative TFBSn ermittelt, die an bestimmten Positionen von mehreren Genen lokalisiert sind. Das führt dazu, dass mehrere Möglichkeiten der Genregulation denkbar sind. Wenn die Zielgene gemeinsam exprimiert werden und der Transkriptionsfaktor koexprimiert wird, ist es denkbar, dass der Faktor die Genexpression induzieren könnte (Singh, 1998). Werden die Gene in bestimmten Geweben nicht exprimiert, der Faktor hingegen schon, könnte das darauf hindeuten, dass der Transkriptionsfaktor die Expression der Gene durch Bindung an die

DNA der Zielgene inhibiert (Sieburth und Meyerowitz, 1997). Wird eine unterschiedliche Expression der Gene untereinander beobachtet und wird der Transkriptionsfaktor entweder exprimiert oder nicht, kann nicht ausgeschlossen werden, dass der Transkriptionsfaktor mit den Genen direkt oder indirekt interagiert. Im Fall von AG aus *Arabidopsis thaliana* ist bekannt, dass die WUS Expression durch AG inhibiert wird. WUS wiederum aktiviert indirekt die Expression des Gens CLV3, welches das Gen CVL1 aktiviert. CVL1 wirkt hingegeben reprimierend auf die WUS-Expression (Bäurle und Laux, 2005, Laux *et al.*, 1996). Die Genregulation ist daher als komplex anzusehen. In der hier vorliegenden Arbeit wurde der Fokus auf die gemeinsame Expression der putativen Zielgene und darüber hinaus auch auf die Koexpression des Transkriptionsfaktors gelegt. Generell lässt sich postulieren, dass sich besonders bei Gemeinsamkeiten in der Expression der Zielgene und des Faktors weiterführende molekularbiologische Analysen anbieten würden, um festzustellen, ob der Transkriptionsfaktor tatsächlich an der Genexpression der putativen Zielgene beteiligt ist.

Die vorliegende Analyse zielte darauf ab, an festgelegten Distanzen relativ zu Genbereichen oder zum intergenischen Bereich Anreicherungen von TFBSn identifizieren zu können. Derartig umfangreiche Analysen sind in der Literatur nicht beschrieben worden. Im intergenischen Bereich, unter der Bedingung, dass das zu der jeweiligen Bindungsstelle gehörende Gen einen Transkriptionsstartpunkt aufweist, war der Transkriptionsfaktor O2 auffällig. Er und seine Zielgene wiesen gemeinsame Expressionen in gekeimten Samen auf. O2 stammt nicht aus *Arabidopsis*, jedoch ist bekannt, dass der Faktor im Endosperm exprimiert wird, dem Nährgewebe für den keimenden Samen (Singh, 1998). Weitere Auffälligkeiten wurden in dem Bereich nicht identifiziert. Es ist auch nicht bekannt, ob die Zielgene von O2 tatsächlich biologisch relevante bZIP-Motive an den entsprechenden Distanzen aufweisen.

Unter der Voraussetzung, dass für das jeweilige Gen der TFBS kein Transkriptionsstartpunkt annotiert ist, war auffällig, dass Gene bei Genevestigator zusammengefasst dargestellt wurden, was sich daher auch in der Genexpression widerspiegelt. Hierbei waren eine Vielzahl Retrotransposons als putative Zielgene des Faktors AtMYB15 bzw. RAV1 vertreten. Retrotransposons werden zunächst transkribiert und translatiert. Danach produziert die reverse Transkriptase eine DNA Kopie der mRNA, welche dann an einer anderen Stelle im Genom durch Integrase-Enzyme eingebaut wird. Daher hat jeder Replikationszyklus von Retrotransposons eine zusätzliche Kopie des Elements zur Folge (Wilhelm and Wilhelm, 2001). Dies würde auch erklären, warum Retrotransposons häufig durch ein einziges Arrayelement bei Genevestigator repräsentiert werden. Durch die hohe Sequenzidentität können die einzelnen Gene durch die Sonden im Array nicht unterschieden werden. *Copia*

Retrotransposons zum Beispiel sind von Long Terminal Repeats (LTRs) umgeben, welche einen Promotor und downstream Elemente beinhalten, die ebenfalls an der Kontrolle der Genexpression beteiligt sind (Wicker und Keller, 2007). Das könnte erklären, warum Bindungsstellen eines Faktors an spezifischen Distanzen gehäuft vorkommen.

In der 5'UTR war bei drei der 10 herausgesuchten Signifikanzen eine gemeinsame Regulation der putativen Zielgene durch ihre entsprechenden Faktoren AtMYB77 bzw. ABI4 zu verzeichnen. In der Literatur ist bislang nicht bekannt, ob die beiden Faktoren auch an DNA-Sequenzen in der 5'UTR von Genen binden können. Generell ist das Vorkommen von Bindungsstellen in der 5'UTR bekannt (Feldbrügge *et al.*, 1994).

In Introns wiesen die herausgesuchten Signifikanzen in wenigen Fällen eine gemeinsame Expression der putativen Zielgene und des entsprechenden Transkriptionsfaktors auf. Bei fünf untersuchten Introns kam es insgesamt in sechs Fällen zu einer gemeinsamen Expression. Interessanterweise war der Faktor AtMYB84 zweimal vertreten, sowie jeweils einmal die Faktoren AGL15, TBP, O2 und TGA1a. Bei den putativen Zielgenen der beiden letztgenannten Faktoren handelte es sich um rezeptorähnliche Kinasen, die wie die Retrotransposons zusammengefasst auf einem Arrayelement bei Genevestigator dargestellt worden sind. Rezeptorähnliche Kinasen (RLKs für receptor like kinase) stellen eine Hauptgenfamilie in Pflanzen dar. Im *Arabidopsis thaliana* Genom gibt es mehr als 600 rezeptorähnliche Kinasen, in Reis sogar mehr als 1100 (Morillo und Tax, 2006). Es wird davon ausgegangen, dass Genduplikation der Hauptfaktor für die große Anzahl an Faktoren dieser Familie ist (Shiu *et al.*, 2004). Unklar ist, ob die Faktoren an der Genregulation beteiligt sind. Im Fall von AtMYB84 ist eine gemeinsame Regulation der Zielgene und des Faktors in ausgereiften Schoten ermittelt worden. Von AtMYB84 ist bislang nicht bekannt, dass Bindungsstellen in Introns auftreten. Gleiches gilt für AGL15 und TBP. Generell sind einige Beispiele für die Lokalisierung von Bindungsstellen von Transkriptionsfaktoren in Introns bekannt (Busch *et al.*, 1999, Tang und Perry, 2003).

Bei den ausgewählten Signifikanzen in den vier untersuchten Exons sind insgesamt in drei Fällen gemeinsame Expressionen von Transkriptionsfaktoren und ihren putativen Zielgenen verzeichnet worden. Die Faktoren sind PIF3, AtMYB84 und AG. Von keinem der Faktoren ist eine Lokalisierung von Bindungsstellen in Exons bekannt. Die Zielgene sind bis auf die für PIF3 recht gemischt, es handelt sich hierbei nicht um gemeinsame Genfamilien oder Retrotransposons. Bei PIF3 gehören einige Gene zu den germinähnlichen Proteinen, die teilweise zusammen auf einem Array bei Genevestigator dargestellt sind. Germinähnliche Proteine (GLPs für Germin-like Proteins) werden durch Multigenfamilien in mehreren

pflanzlichen Spezies kodiert und spielen z.B. bei der Pathogenabwehr eine Rolle (Zimmermann *et al.*, 2006). Die Funktion der meisten GLPs bei Pflanzen ist jedoch meist unbekannt (Zimmermann *et al.*, 2006). Es ist anzunehmen, dass sich die Anzahl der Proteine einer Familie durch Genduplikation im Laufe der Zeit vergrößert hat und dass die identischen Nukleotide nun bei der hier vorliegenden Analyse zu einer Feststellung von einer Signifikanz geführt haben. Interessant bei den GLPs ist, dass sie oft robust gegen Denaturierung und Degradierung durch Hitze, Proteinasen und Detergenzien sind (Vallelian-Bindschedler et al., 1998). Generell ist sehr wenig über Bindungsstellen von Transkriptionsfaktoren in Exons bekannt. Dieses Gebiet gilt es also noch zu erforschen.

In der 3'UTR sind bei drei der 10 Signifikanzen tatsächlich Übereinstimmungen bei der Regulation der putativen Zielgene und dem entsprechenden Transkriptionsfaktor festgestellt worden. Die Faktoren sind ATHB5 (in zwei Fällen) und CDC5. Die Zielgene von ATHB5 überlappen nicht. Es ist bislang nicht bekannt, dass die beiden Faktoren Bindungsstellen in der 3'UTR von Genen aufweisen. Das prinzipielle Auftreten von Bindungsstellen in der 3'UTR jedoch ist in der Literatur beschrieben worden (Banks *et al.*, 2000, Hyder *et al.*, 2000).

Prinzipiell deutet vieles darauf hin, dass die in dieser Arbeit entwickelte Vorgehensweise sinnvoll eingesetzt werden kann, um mögliche Zielgene von Transkriptionsfaktoren vorherzusagen und die Expression von beiden zu untersuchen. Es bietet sich an, die vorhergesagten Ergebnisse mit molekularbiologischen Methoden zu verifizieren.

5. Zusammenfassung

Für bioinformatische Analysen ist das *Arabidopsis thaliana* Genom sehr gut geeignet, da es durchsequenziert ist, und die Pflanze zu den Modellorganismen gehört.
Im Rahmen dieser Arbeit wurden Verteilungsanalysen von putativen Transkriptionsfaktor-Bindungsstellen (TFBSn) in *Arabidopsis* durchgeführt.
Zunächst wurden die fünf Chromosomen analysiert. Gene sind im Durchschnitt 2246 Nukleotide lang und weisen neben einer 5'UTR und 3'UTR im Schnitt etwa vier Exons und etwa fünf Introns auf. Als das erste Exon wurde ein Exon definiert, wenn es mit einem Translationsstartcodon anfängt, als das letzte Exon wurde ein Exon definiert, wenn es mit einem Translationsstoppcodon endet. Die 5'UTR hat eine mittlere Länge von 119, das erste Exon 692, die weiteren vier 259 bis 192, die fünf Introns 248 bis 136 und die 3'UTR 223 Nukleotiden.
Der AT-Gehalt beträgt im Durchschnitt 64,2 %, verteilt auf den intergenischen Bereich (68,47 %), die 5'UTR (61,84 %), die Introns (67,59 %), die Exons (56,39 %) und die 3'UTR (66,47 %). Bei den untersuchten verschiedenen Introns und Exons innerhalb von Genen bleibt der AT-Gehalt jeweils etwa gleich.
Die einzelnen Bereiche innerhalb von Genen wurden auf konservierte Sequenzmotive und Nukleotide hin analysiert. Die 5'UTR beginnt bevorzugt mit AT und endet bevorzugt mit AAAA. Introns weisen zu Beginn präferenziell die Sequenz GTAAGT und am Ende bevorzugt T(G)(C/T)AG auf. Das erste Exon eines Gens zeigt eine Konservierung für ATG(G) am Beginn, während am Ende des letzten Exons vorwiegend T(A/G)(A/G) vorkommt. Exons innerhalb von Genen beginnen bevorzugt mit GT und enden mit A(G/A). Die 3'UTR weist keine bevorzugten Nukleotide am Beginn auf, generell ist sie T- und A-reich. Kurz vor dem Ende der 3'UTR ist eine deutlichere Präferenz für das Nukleotid T zu verzeichnen.

Im Rahmen der Arbeit wurden Verteilungsanalysen von TFBSn im *Arabidopsis thaliana* Genom mit Matrizen durchgeführt. Matrizen, die eine Sammlung von TFBSn für jeweils einen Transkriptionsfaktor beinhalten, enthalten eine Kernsequenz sowie flankierende Nukleotide. Insgesamt wurden 42 pflanzliche Matrizen für die Analysen eingesetzt. Mit diesen Matrizen wurde eine genomweite Suche nach putativen TFBSn durchgeführt. Um die erhaltenen Ergebnisse einschätzen zu können, wurden Hintergrundmodelle verwendet. Einerseits wurden Zufallspositionen generiert. Diese Daten lieferten Werte unter der Annahme der zufälligen Verteilung von TFBSn. Des Weiteren wurden Zufallsmatrizen für die pflanzlichen Matrizen generiert, unter der Bedingung, dass sie den gleichen AT-Gehalt sowie die gleiche Anzahl an in die jeweilige pflanzliche Matrize eingehenden Sequenzen aufweisen. Insgesamt

wurden 42 Zufallsmatrizen für die pflanzlichen Matrizen generiert. Auf diese Weise konnte eine Einschätzung der Verteilung aufgrund des AT-Gehalts der pflanzlichen Matrize durchgeführt werden. Nichtpflanzliche Matrizen sind wie pflanzliche Matrizen aufgebaut. Sie weisen eine Kernsequenz sowie flankierende, variable Nukleotide auf. Zur weiteren Einschätzung der Verteilungen der TFBSn der pflanzlichen Matrizen wurden nichtpflanzliche Matrizen eingesetzt, um den Einfluss einer Kernsequenz auf die Verteilung zu untersuchen. Analog zu der Verfahrensweise bei den pflanzlichen Matrizen wurden auch für die nichtpflanzlichen Matrizen Zufallsmatrizen generiert und eingesetzt.

Analysiert wurde die Verteilung von putativen TFBSn relativ zum nächstgelegenen Translationsstartpunkt. Diese Analyse wurde zunächst in 50 bp Schritten durchgeführt. Drei Verteilungstypen konnten identifiziert werden. Eine upstreamlastige Verteilung weist eine Anhäufung von Bindungsstellen im upstream Bereich von Genen auf, eine downstreamlastige Verteilung eine Anreicherung von Bindungsstellen im downstream Bereich. Eine indifferente Verteilung weist keine Bevorzugung für einen dieser Bereiche auf. Zufallspositionen weisen eine indifferente Verteilung auf. Bindungsstellen, die mit pflanzlichen Matrizen nachgewiesen werden, sind zu 21,42 % upstreamlastig, zu 35,71 % downstreamlastig und zu 42,86 % indifferent verteilt. Die Bindungsstellen der Zufallsmatrizen der pflanzlichen Matrizen sind vorwiegend downstreamlastig verteilt (71,43 %). Eine upstreamlastige Verteilung sowie eine indifferente kommen gleich häufig vor (jeweils 14,29 %). Die Art der Verteilung hängt nicht allein vom AT-Gehalt einer Matrize ab. Bindungsstellen nichtpflanzlicher Matrizen sind vorwiegend downstreamlastig verteilt (59,26 %). Die upstreamlastige Verteilung ist die Ausnahme (3,7 %), während 37,04 % der Bindungsstellen der Matrizen eine indifferente Verteilung aufweisen. Im Vergleich dazu sind die Bindungsstellen der Zufallsmatrizen der nichtpflanzlichen Matrizen ebenfalls vorwiegend downstreamlastig verteilt (85,16 %). Eine upstreamlastige Verteilung bzw. eine indifferente Verteilung kommen gleich häufig vor (jeweils 7,41 %). Die Art der Verteilung hängt auch hier nicht vom AT-Gehalt der Matrize allein ab. Der Vergleich zwischen der Verteilung von TFBSn pflanzlicher Matrizen und nichtpflanzlicher Matrizen zeigt Gemeinsamkeiten, aber auch Unterschiede. Daher kann auch eine Kernsequenz nicht der alleinige Parameter sein, der die Art der Verteilung determiniert. Es wird davon ausgegangen, dass die Bindungsstellen pflanzlicher Matrizen reale Bindungsstellen beinhalten können.

Während dieser Arbeit wurde das Vorkommen der putativen Bindungsstellen im intergenischen Bereich sowie zunächst in der 5'UTR, in Introns, Exons und der 3'UTR untersucht. Anschließend wurden die Werte für die erste 5'UTR, die ersten fünf Introns, die ersten vier Exons und die erste 3'UTR analysiert. Die meisten Bindungsstellen kommen im

Exon vor, gefolgt vom intergenischen Bereich. Deutlich weniger putative Bindungsstellen treten in Introns auf. Sehr wenige Bindungsstellen sind in der 3'UTR und in der 5'UTR zu verzeichnen. Bei der Analyse der ersten vier Exons treten die Bindungsstellen gehäuft im ersten Exon auf und sind in den anderen Exons entsprechend geringer. Dies beruht wahrscheinlich darauf, dass die Länge des ersten Exons deutlich größer ist als die der übrigen Exons. Die prinzipielle Feststellung, dass Bindungsstellen vorwiegend in Exons auftreten, kann mit einem fehlenden Selektionsdruck zusammenhängen. Die Anhäufung von Bindungsstellen im intergenischen Bereich kann darauf beruhen, dass die mit den Matrizen detektierten Bindungsstellen eine Vielzahl von realen Bindungsstellen enthalten können.
Es wurde die Anzahl der putativen Bindungsstellen jeder pflanzlichen Matrize mit der der Zufallspositionen verglichen. Hierbei wurden Abweichungsfaktoren der Bindungsstellen der pflanzlichen Matrizen von denen der Zufallspositionen gebildet. Es kommt bei 40,48 % der Matrizen im intergenischen Bereich zu positiven Abweichungsfaktoren von den Zufallspositionen. 26,19 % der TFBSn der Matrizen zeigen in der ersten 5'UTR positive Abweichungsfaktoren. Bei den untersuchten ersten fünf Introns bleibt die Abweichungsrichtung der Bindungsstellen der Matrizen vorwiegend gleich. Gleiches gilt für die untersuchten ersten vier Exons. In der 3'UTR kommt es bei 14,29 % der TFBSn pflanzlichen Matrizen zu positiven Abweichungsfaktoren von den Zufallspositionen.

Es wurde die Abweichung von putativen TFBSn im intergenischen Bereich, untersucht in 1 bp Schritten, bestimmt. Hierzu wurde jeweils ein Abweichungsfaktor von den Zufallspositionen gebildet. Ein positiver Abweichungsfaktor bedeutet, dass mehr TFBSn vorkommen als Zufallspositionen, ein negativer Abweichungsfaktor belegt, dass weniger TFBSn als Zufallspositionen auftreten. Bei allen vier Matrizentypen (pflanzliche Matrizen, Zufallsmatrizen für die pflanzlichen Matrizen, nichtpflanzliche Matrizen und Zufallsmatrizen für die nichtpflanzlichen Matrizen) konnten im upstream-Bereich negative Abweichungsfaktoren identifiziert werden, die mit abnehmender Distanz zum Gen zunehmen. Dieser Bereich beträgt im Durchschnitt -405 bp bis -3 bp relativ zum Transkriptionsstart für Gene mit annotiertem Transkriptionsstart, und -358 bp bis -12 bp relativ zum Translationsstart für Gene ohne Transkriptionsstart. Das deutet auf eine Gegenselektion von zufälligen und nicht regulatorischen DNA-Sequenzen hin. In diesem Bereich könnten die meisten regulatorischen DNA-Sequenzen liegen, die für die Regulation der Transkription eines Gens wichtig sind.
Direkte Überlagerungsgrafiken, die sowohl die Abweichungsfaktoren der pflanzlichen Matrize als auch die ihrer entsprechenden Zufallsmatrize beinhalten, zeigen bei den pflanzlichen Matrizen Distanzbereiche relativ zum nächstgelegenen Gen, die deutlich mehr pflanzliche Bindungsstellen aufweisen als dies bei den entsprechenden Zufallsmatrizen der Fall ist.

Diese Grafiken wurden für den intergenischen Bereich erstellt. Derartige Distanzbereiche treten bei den nichtpflanzlichen Matrizen nicht auf. Das deutet darauf hin, dass es sich um eine Anhäufung von realen Bindungsstellen in diesen Bereichen handeln könnte. Aufgefallen sind diese Distanzbereiche vorwiegend bei Matrizen, deren dazugehörige Transkriptionsfaktoren zu den Klassen bZIP, bHLH und TCP gehören. Einzelne Distanzen, an denen deutlich mehr putative Bindungsstellen der pflanzlichen Matrizen auftreten als bei ihren entsprechenden Zufallsmatrizen können auch festgestellt werden. Diese Art der Auffälligkeit ist auch bei den nichtpflanzlichen Matrizen aufzufinden.

Es wurde ein statistisches Kriterium, ein P-Wert, eingesetzt, der ein Maß für die Signifikanz der festgestellten Bindungsstellen in einzelnen Distanzbereichen darstellt. Eine putative Bindungsstelle eines Transkriptionsfaktors wird einem Gen zugeordnet und kann an der Regulation des Gens beteiligt sein. Wird von einer Beteiligung ausgegangen, stellt dieses Gen ein Zielgen des Transkriptionsfaktors dar. Der P-Wert gibt an, inwieweit das Vorkommen der TFBS aufgrund des Zufalls zu erwarten ist. Die Auswertung wurde auf die pflanzlichen Matrizen beschränkt. Die jeweils 10 signifikantesten P-Werte im intergenischen Bereich, in der ersten 5'UTR, den ersten fünf Introns, den ersten vier Exons und der ersten 3'UTR wurden identifiziert. Die entsprechenden Gene der TFBSn wurden jeweils bei Genevestigator eingesetzt, um Genexpressionsanalysen zu erhalten. Ziel war es, gemeinsame Expressionen der putativen Zielgene des zu der Matrize gehörenden Faktors mit dem Faktor selbst zu erhalten.

Bei Bindungsstellen in der ersten 5'UTR (Distanz 70 Nukleotide) ist eine Übereinstimmung in der Genexpression des Faktors AtMYB77 und seiner potenziellen Zielgene zu verzeichnen. Des Weiteren stimmen die Expressionsdaten von ABI4 und seiner möglichen Zielgene (Distanz 96 bp sowie Distanz 32 bp) überein. Die Zielgene unterscheiden sich bei den identifizierten Distanzen. Betreffend der Analysen für das dritte Intron stimmen die Expressionsdaten von AtMYB84 (Distanz 7) mit denen der putativen Zielgenen überein. Die Expressionsdaten von AGL15 (Distanz 3) bzw. AtMYB84 (Distanz 7) bezüglich der TFBSn im fünften Intron erbrachten eine Übereinstimmung mit den jeweiligen Zielgenen. Hinsichtlich der TFBSn im zweiten Exon zeigen die putativen Zielgene von PIF3 (Distanz 227) gleiche Expressionsdaten wie der Faktor. Die putativen Zielgene von AtMYB84 weisen betreffend der Analysen im dritten Exon (Distanz 3) gleiche Expressionsdaten wie der Faktor selbst auf. Bei den Analysen hinsichtlich der TFBSn im vierten Exon tritt eine übereinstimmende Genexpression der putativen Zielgene des Faktors AG (Distanz 4) auf. Die Expressionsdaten des Faktors ATHB5 (Distanz 6 bzw. 208) bezüglich der 3'UTR zeigen gleiche Daten wie die der entsprechenden jeweiligen putativen Zielgene. Beim intergenischen Bereich wurden zwei Kriterien angewendet. Einerseits wurden Bindungsstellen untersucht, deren dazugehörige Gene einen Transkriptionsstart aufweisen und andererseits diese, deren dazugehörige Gene

keinen Transkriptionsstart, aber einen Translationsstart besitzen. Unter der Bedingung, dass der Transkriptionsstart vorhanden ist, kann bei den putativen Zielgenen von O2 und den entsprechenden putativen Zielgenen (Distanz -79) eine Übereinstimmung in den Expressionsdaten festgestellt werden. Unter der Voraussetzung, dass kein Transkriptionsstartpunkt, aber ein Translationsstartpunkt annotiert ist, ist bei AtMYB15 (Distanz -56), RAV1 (Distanz -687) und AGL15 (Distanz -8) eine gemeinsame Expression der jeweiligen Faktoren und ihren putativen Zielgenen zu verzeichnen.
Insgesamt treten gemeinsame Expressionen von TFBSn und dem Transkriptionsfaktor nicht nur bei Bindungsstellen im intergenischen Bereich sondern auch bei Bindungsstellen in einzelnen Bereichen innerhalb von Genen auf.

6. Literaturverzeichnis

Adachi, K., Toyota, M., Sasaki, Y., Yamashita, T., Ishida, S., Ohe-Toyota, M., Maruyama, R., Hinoda, Y., Saito, T., Imai, K., Kudo, R., Tokino, T. (2004) Identification of SCN3B as a novel p53-inducible proapoptotic gene, Oncogene 23: 7791-8

Alexandrov, N.N., Troukhan, M.E., Brover, V.V., Tatarinova, T., Flavell, R.B., Feldmann, K.A. (2006) Features of Arabidopsis genes and genome discovered using full-length cDNAs, Plant Mol. Biol. 60: 69-85

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. (1990) Basic Local Alignment Search Tool. J. Mol. Biol. 215:403–410

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25:3389-402

Apfel, R., Benbrook, D., Lernhardt, E., Ortiz, M.A., Salbert, G., Pfahl, M. (1994) A novel orphan receptor specific for a subset of thyroid hormone-responsive elements and its interaction with the retinoid/thyroid hormone receptor subfamily, Mol. Cell Biol. 14: 7025-35

Arabidopsis Genome Initiative. (2000), Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana,* Nature 408: 796-815.

Banks, G.C., Li, Y., Reeves, R. (2000) Differential *in vivo* modifications of the HMGI(Y) nonhistone chromatin proteins modulate nucleosome and DNA interactions, Biochemistry 39: 8333-46

Bartsch, H.-J. (1994) Taschenbuch mathematischer Formeln, 16. Auflage, Fachbuchverlag Leipzig-Köln

Bastola, D.R., Pethe, V.V., Winicov, I. (1998) Alfin1, a novel zinc-finger protein in alfalfa roots that binds to promoter elements in the salt-inducible MsPRP2 gene, Plant Mol. Biol. 38: 1123-35

Bäurle, I., Laux, T. (2005) Regulation of WUSCHEL transcription in the stem cell niche of the *Arabidopsis* shoot meristem. Plant Cell. 17:2271-80.

Behnen, K., Neuhaus, G. (1995) Grundkurs Stochastik, Eine integrierte Einführung in Wahrscheinlichkeitstheorie und Mathematische Stochastik, 3. Auflage, B.G.Teubner Stuttgart

Berendzen, K.W., Stüber, K., Harter, K., Wanke, D. (2006) *Cis*-motifs upstream of the transcription and translation initiation sites are effectively revealed by their positional disequilibrium in eukaryote genomes using frequency distribution curves, BMC Bioinformatics 7:522

Blancafort, P., Segal, D.J., Barbas, C.F. 3rd. (2004) Designing transcription factor architectures for drug discovery, Mol. Pharmacol. 66: 1361-71

Brewer, P.B., Howles, P.A., Dorian, K., Griffith, M.E., Ishida, T., Kaplan-Levy, R.N., Kilinc, A., Smyth, D.R. (2004) PETAL LOSS, a trihelix transcription factor gene, regulates perianth architecture in the Arabidopsis flower, Development 131: 4035-45

Brooks, A.R., Blackhart, B.D., Haubold, K., Levy-Wilson, B. (1991) Characterization of tissue-specific enhancer elements in the second intron of the human apolipoprotein B gene, J. Biol. Chem. 266: 7848-59

Busch, M.A., Bomblies, K., Weigel, D. (1999) Activation of a floral homeotic gene in *Arabidopsis*, Science 285: 585-7

Bülow, L., Steffens, N.O., Galuschka, C., Schindler, M., Hehl, R. (2006) AthaMap: from in silico data to real transcription factor binding sites, In Silico Biol. 6: 243-52

Byrne, M.E. (2006) Shoot meristem function and leaf polarity: the role of class III HD-ZIP genes, PLoS Genet. 2: e89

Cavener, D.R. (1987) Comparison of the consensus sequence flanking translational start sites in Drosophila and vertebrates, Nucleic Acids Res., 15: 1351-1361

Chan, R.L., Gago, G.M., Palena, C.M., Gonzalez, D.H. (1998) Homeoboxes in plant development, Biochim. Biophys. Acta.: 1442: 1-19

Chew, L.J., Huang, F., Boutin, J.M., Gallo, V. (1999) Identification of nuclear orphan receptors as regulators of expression of a neurotransmitter receptor gene, J. Biol. Chem. 274: 29366-75

Chinnusamy, V., Ohta, M., Kanrar, S., Lee, B.H., Hong, X., Agarwal, M., Zhu, J.K. (2003) ICE1: a regulator of cold-induced transcriptome and freezing tolerance in *Arabidopsis*, Genes Dev. 17: 1043-54

Choi, H., Hong, J., Ha, J., Kang, J., Kim, S.Y. (2000) ABFs, a family of ABA-responsive element binding factors, J. Biol. Chem. 275: 1723-30

Chung, B.Y., Simons, C., Firth, A.E., Brown, C.M., Hellens, R.P. (2006) Effect of 5'UTR introns on gene expression in Arabidopsis thaliana, BMC Genomics 7: 120

Crooks, G.E., Hon, G., Chandonia, J.M., Brenner, S.E. (2004) WebLogo: a sequence logo generator, Genome Res. 14: 1188-90

Davuluri, R.V., Sun, H., Palaniswamy, S.K., Matthews, N., Molina, C., Kurtz, M. and Grotewold, E. (2003) AGRIS: *Arabidopsis* Gene Regulatory Information Server, an information resource of *Arabidopsis* *cis*-regulatory elements and transcription factors. BMC Bioinformatics: 4, 25

Day, D.A., Tuite, M.F. (1998) Post-transcriptional gene regulatory mechanisms in eukaryotes: an overview, J. Endocrinol. 157: 361-71

Delseny, M. (2003) Towards an accurate sequence of the rice genome, Curr. Opin. Plant Biol. 6: 101-5

Demura, T., Fukuda, H. (2007) Transcriptional regulation in wood formation, Trends Plant Sci. 12: 64-70

de Pater, S., Greco, V., Pham, K., Memelink, J., Kijne, J. (1996) Characterization of a zinc-dependent transcriptional activator from *Arabidopsis*, Nucleic Acids Res. 24: 4624-31

Doan, L.L., Porter, S.D., Duan, Z., Flubacher, M.M., Montoya, D., Tsichlis, P.N., Horwitz, M., Gilks, C.B., Grimes, H.L. (2004) Targeted transcriptional repression of Gfi1 by GFI1 and GFI1B in lymphoid cells, Nucleic Acids Res. 32: 2508-19

Duek, P.D., Fankhauser, C. (2005) bHLH class transcription factors take centre stage in phytochrome signalling, Trends Plant Sci. 10: 51-4

Duval, M., Hsieh, T.F., Kim, S.Y., Thomas, T.L. (2002) Molecular characterization of AtNAM: a member of the Arabidopsis NAC domain superfamily, Plant Mol. Biol. 50: 237-48

Eulgem, T., Rushton, P.J., Robatzek, S., Somssich, I.E. (2000) The WRKY superfamily of plant transcription factors, Trends Plant Sci. 5: 199-206

Farnham, P.J., Means, A.L. (1990) Sequences downstream of the transcription initiation site modulate the activity of the murine dihydrofolate reductase promoter, Mol. Cell Biol. 10: 1390-8

Feldbrügge, M., Sprenger, M., Dinkelbach, M., Yazaki, K., Harter, K., Weisshaar, B. (1994) Functional analysis of a light-responsive plant bZIP transcriptional regulator, Plant Cell 6: 1607-21

Ferrario, S., Immink, R.G., Shchennikova, A., Busscher-Lange, J., Angenent, G.C. (2003) The MADS box gene FBP2 is required for SEPALLATA function in petunia, Plant Cell 15: 914-25

Galuschka, C., Schindler, M., Bülow, L., Hehl, R. (2007) AthaMap web tools for the analysis and identification of co-regulated genes, Nucleic Acids Res. 35: D857-62

Grennan, A.K. (2006) Genevestigator. Facilitating web-based gene-expression analysis, Plant Physiol. 14:1164-6

Grotewold, E., Drummond, B.J., Bowen, B., Peterson, T. (1994) The myb-homologous P gene controls phlobaphene pigmentation in maize floral organs by directly activating a flavonoid biosynthetic gene subset, Cell. 76:543-53

Gubler, F., Raventos, D., Keys, M., Watts, R., Mundy, J., Jacobsen, J.V. (1999) Target genes and regulatory domains of the GAMYB transcriptional activator in cereal aleurone, Plant J.: 17:1-9

Gutterson, N., Reuber, T.L. (2004) Regulation of disease resistance pathways by AP2/ERF transcription factors, Curr. Opin. Plant Biol. 7: 465-71

Haas, B.J., Wortman, J.R., Ronning, C.M., Hannick, L.I., Smith, R.K., Maiti, R., Chan, A.P., Yu, C., Farzad, M., Wu, D., White, O., Town, C.D. (2005) Complete reannotation of the Arabidopsis genome: methods, tools, protocols and the final release, BMC Biol.: 3: 7

Hannenhalli, S., Levy, S. (2002) Predicting transcription factor synergism, Nucleic Acids Res. 30: 4278-4284

Hertz, G. Z., Stormo, G. D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences, Bioinformatics 15: 563-577

Hirayama, T., Shinozaki, K. (1996) A cdc5+ homolog of a higher plant, *Arabidopsis thaliana*, Proc. Natl. Acad. Sci. U S A. 93:13371-6

Huang, H., Mizukami, Y., Hu, Y., Ma, H. (1993) Isolation and characterization of the binding sequences for the product of the *Arabidopsis* floral homeotic gene AGAMOUS, Nucleic Acids Res. 21: 4769-76

Huang, H., Tudor, M., Weiss, C.A., Hu, Y., Ma, H. (1995) The *Arabidopsis* MADS-box gene AGL3 is widely expressed and encodes a sequence-specific DNA-binding protein, Plant Mol. Biol. 28: 549-67

Hyder, S.M., Nawaz, Z., Chiappetta, C., Stancel, G.M. (2000) Identification of functional estrogen response elements in the gene coding for the potent angiogenic factor vascular endothelial growth factor, Cancer Res. 60: 3183-90

Ingram, G.C., Waites, R. (2006) Keeping it together: co-ordinating plant growth, Curr. Opin. Plant Biol. 9: 12-20

Istrail, S., Davidson, E.H. (2005) Logic functions of the genomic cis-regulatory code, Proc Natl. Acad. Sci. U S A 102: 4954-9

Izawa, T., Foster, R., Chua, N.H. (1993) Plant bZIP protein DNA binding specificity, J. Mol. Biol 230: 1131-44

Jack, T. (2001) Plant development going MADS, Plant Mol. Biol., 46: 515-520

Jakoby, M., Weisshaar, B., Dröge-Laser, W., Vicente-Carbajosa, J., Tiedemann, J, Kroj, T., Parcy, F. (2002) bZIP transcription factors in *Arabidopsis*, Trends Plant Sci., 7: 106-111

Jeong, Y.M., Mun, J.H., Lee, I., Woo, J.C., Hong, C.B., Kim, S.G. (2006) Distinct roles of the first introns on the expression of Arabidopsis profilin gene family members, Plant Physiol. 140: 196-209

Jin, H., Martin, C. (1999) Multifunction and diversity within the plant MYB-gene family, Plant Mol. Biol., 41: 577-585

Johannesson, H., Wang, Y., Engstrom, P. (2001) DNA-binding and dimerization preferences of *Arabidopsis* homeodomain-leucine zipper transcription factors *in vitro*, Plant Mol. Biol. 45:63-73

Kagaya, Y., Ohmiya, K., Hattori, T. (1999) RAV1, a novel DNA-binding protein, binds to bipartite recognition sequence through two distinct DNA-binding domains uniquely found in higher plants, Nucleic Acids Res. 27: 470-8

Kawaguchi, R., Bailey-Serres, J. (2005) mRNA sequence features that contribute to translational regulation in Arabidopsis, Nucleic Acids Res. 33: 955-65

Knippers, R. (1997) Molekulare Genetik, 7. Auflage, Georg Thieme Verlag Stuttgart New York

Kornblihtt, A.R., de la Mata, M., Fededa, J.P., Munoz, M.J., Nogues, G. (2004) Multiple links between transcription and splicing, RNA 10: 1489-98

Kosugi, S., Ohashi, Y. (2000) Cloning and DNA-binding properties of a tobacco Ethylene-Insensitive3 (EIN3) homolog, Nucleic Acids Res. 28: 960-7

Kosugi, S., Ohashi, Y. (2002) DNA binding and dimerization specificity and potential targets for the TCP protein family, Plant J. 30: 337-48

Kozaki, A., Hake, S., Colasanti, J. (2004) The maize ID1 flowering time regulator is a zinc finger protein with novel DNA binding properties, Nucleic Acids Res. 32: 1710-20

Krusell, L., Rasmussen, I., Gausing, K. (1997) DNA binding sites recognised *in vitro* by a knotted class 1 homeodomain protein encoded by the hooded gene, k, in barley (*Hordeum vulgare*), FEBS Lett. 408: 25-9

Laux, T., Mayer, K.F., Berger, J., Jürgens, G. (1996) The WUSCHEL gene is required for shoot and floral meristem integrity in *Arabidopsis* , Development 122:87-96.

Liu, L., White, M.J., MacRae, T.H. (1999) Transcription factors and their genes in higher plants functional domains, evolution and regulation, Eur. J. Biochem. 262: 247-57

Lozán, J.L., Kausch, H. (2004) Angewandte Statistik für Naturwissenschaftler, 3. Auflage, Parey Verlag, Hamburg

Martinez-Garcia, J.F., Huq, E., Quail, P.H. (2000) Direct targeting of light signals to a promoter element-bound transcription factor, Science 288: 859-63

Martinez-Garcia, J.F., Moyano, E., Alcocer, M.J., Martin, C. (1998) Two bZIP proteins from *Antirrhinum* flowers preferentially bind a hybrid C-box/G-box motif and help to define a new sub-family of bZIP transcription factors, Plant J. 13: 489-505

Matys, V., Fricke, E., Geffers, R., Gössling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V., Kloos, D.U., Land, S., Lewicki-Potapov, B., Michael, H., Münch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S., Wingender, E.. (2003), TRANSFAC: transcriptional regulation, from patterns to profiles, Nucleic Acids Res. 31: 374-8

Meli, M., Albert-Fournier, B., Maurel, M.C. (2001) Recent findings in the modern RNA world, Int. Microbiol. 4:5-11

Mellor, J. (2006) It takes a PHD to read the histone code, Cell 126: 22-4

Merriam, L.C., Chess, A. (2007) *cis*-Regulatory elements within the odorant receptor coding region, Cell. 131:844-6

Mizuno, M., Kanehisa, M. (1994) Distribution profiles of GC content around the translation initiation site in different species, FEBS Lett. 352: 7-10

Molina, C., Grotewold, E. (2005) Genome wide analysis of Arabidopsis core promoters, BMC Genomics 6: 25

Morillo, S.A., Tax, F.E. (2006) Functional analysis of receptor-like kinases in monocots and dicots, Curr. Opin. Plant Biol. 9:460-469

Muszynski, M.G., Dam, T., Li, B., Shirbroun, D.M., Hou, Z., Bruggemann, E., Archibald, R., Ananiev, E.V., Danilevskaya, O.N. (2006) delayed flowering1 Encodes a basic leucine zipper protein that mediates floral inductive signals at the shoot apex in maize, Plant Physiol. 142:1523-1536

Naito, T., Kiba, T., Koizumi, N., Yamashino, T., Mizuno, T. (2007) Characterization of a unique GATA family gene that responds to both light and cytokinin in Arabidopsis *thaliana*, Biosci. Biotechnol. Biochem. 71: 1557-60

Navaud, O., Dabos, P., Carnus, E., Tremousaygue, D., Hervé, C. (2007) TCP transcription factors predate the emergence of land plants, J. Mol. Evol. 65: 23-33

Niu, X., Helentjaris, T., Bate, N.J. (2002) Maize ABI4 binds coupling element1 in abscisic acid and sugar response genes, Plant Cell 14: 2565-75

Nole-Wilson, S., Krizek, B.A. (2000) DNA binding properties of the *Arabidopsis* floral development protein AINTEGUMENTA, Nucleic Acids Res. 28: 4076-82

O'Connor, T.R., Dyreson, C., Wyrick, J.J. (2005) Athena: a resource for rapid visualization and systematic analysis of *Arabidopsis* promoter sequences, Bioinformatics 21:4411-4413

Ouwerkerk, P.B., Trimborn, T.O., Hilliou, F., Memelink, J. (1999) Nuclear factors GT-1 and 3AF1 interact with multiple sequences within the promoter of the Tdc gene from *Madagascar periwinkle*: GT-1 is involved in UV light-induced expression, Mol. Gen. Genet. 261: 610-22

Palena, C.M., Gonzalez, D.H., Chan, R.L. (1999) A monomer-dimer equilibrium modulates the interaction of the sunflower homeodomain leucine-zipper protein Hahb-4 with DNA, Biochem. J. 341: 81-7

Pedersen, A.G., Baldi, P., Chauvin, Y., Brunak, S. (1999) The biology of eukaryotic promoter prediction, Comput Chem. 23:191-207

Ramsay, N.A., Glover, B.J. (2005) MYB-bHLH-WD40 protein complex and the evolution of cellular diversity, Trends Plant Sci. 10: 63-70

Reiser, L., Sánchez-Baracaldo, P., Hake, S. (2000) Knots in the family tree: evolutionary relationships and functions of knox homeobox genes, Plant Mol. Biol. 42: 151-66

Reményi, A., Schöler, H.R., Wilmanns, M. (2004) Combinatorial control of gene expression, Nat. Struct. Mol. Biol. 11: 812-5

Reyes, J.C., Muro-Pastor, M.I., Florencio, F.J. (2004) The GATA family of transcription factors in Arabidopsis and rice, Plant Physiol. 134: 1718-32

Riechmann, J.L., Heard, J., Martin, G., Reuber, L., Jiang, C., Keddie, J., Adam, L., Pineda, O., Ratcliffe, O.J., Samaha, R.R., Creelman, R., Pilgrim, M., Broun, P., Zhang, J.Z., Ghandehari, D., Sherman, B.K., Yu, G. (2000) *Arabidopsis* transcription factors: genome-wide comparative analysis among eukaryotes, Science 290: 2105-10

Riechmann, J.L., Ito, T., Meyerowitz, E.M. (1999) Non-AUG initiation of AGAMOUS mRNA translation in Arabidopsis thaliana, Mol. Cell Biol. 19: 8505-12

Riechmann, J.L., Meyerowitz, E.M. (1998) The AP2/EREBP family of plant transcription factors, Biol. Chem., 379: 633-646

Rijpkema, A.S., Gerats, T., Vandenbussche, M. (2007) Evolutionary complexity of MADS complexes, Curr. Opin. Plant Biol. 10: 32-8

Romero, I., Fuertes, A., Benito, M.J., Malpica, J.M., Leyva, A., Paz-Ares, J. (1998) More than 80R2R3-MYB regulatory genes in the genome of *Arabidopsis thaliana*, Plant J. 14: 273-84

Sakamoto, T., Kamiya, N., Ueguchi-Tanaka, M., Iwahori, S., Matsuoka, M. (2001) KNOX homeodomain protein directly suppresses the expression of a gibberellin biosynthetic gene in the tobacco shoot apical meristem, Genes Dev. 15: 581-90

Schindler, U., Beckmann, H., Cashmore, A.R. (1992) TGA1 and G-box binding factors: two distinct classes of *Arabidopsis* leucine zipper proteins compete for the G-box-like element TGACGTGG, Plant Cell 4: 1309-19

Schneider, T.D., Stephens, R.M. (1990) Sequence logos: a new way to display consensus sequences, Nucleic Acids Res. 18: 6097-100

Seki, M., Ishida, J., Narusaka, M., Fujita, M., Nanjo, T., Umezawa, T., Kamiya, A., Nakajima, M., Enju, A., Sakurai, T., Satou, M., Akiyama, K., Yamaguchi-Shinozaki, K., Carninci, P., Kawai, J., Hayashizaki, Y., Shinozaki, K. (2002) Monitoring the expression pattern of around 7,000 *Arabidopsis* genes under ABA treatments using a full-length cDNA microarray, Funct. Integr. Genomics 2:282-91

Sessa, G., Morelli, G., Ruberti, I. (1993) The Athb-1 and -2 HD-Zip domains homodimerize forming complexes of different DNA binding specificities, EMBO J. 12: 3507-17

Sessa, G., Steindler, C., Morelli, G., Ruberti, I. (1998) The *Arabidopsis* Athb-8, -9 and -14 genes are members of a small gene family coding for highly related HD-ZIP proteins, Plant Mol. Biol. 38: 609-22

Shahmuradov, I.A., Gammerman, A.J., Hancock, J.M., Bramley, P.M., Solovyev, V.V. (2003) PlantProm: a database of plant promoter sequences, Nucleic Acids Res. 31: 114-7

Shiu, S.H., Karlowski, W.M., Pan, R., Tzeng, Y.H., Mayer, K.F., Li, W.H. (2004) Comparative analysis of the receptor-like kinase family in Arabidopsis and rice, Plant Cell 16:1220-1234

Sieburth, L.E., Meyerowitz, E.M. (1997) Molecular dissection of the AGAMOUS control region shows that cis elements for spatial regulation are located intragenically, Plant Cell 9: 355-65

Singh, K.B. (1998) Transcriptional regulation in plants: the importance of combinatorial control, Plant Physiol. 118: 1111-20

Solano, R., Nieto, C., Avila, J., Canas, L., Diaz, I., Paz-Ares, J. (1995) Dual DNA binding specificity of a petal epidermis-specific MYB transcription factor (MYB.Ph3) from *Petunia hybrida*, EMBO J. 14:1773-84

Steffens, N.O., Galuschka, C., Schindler, M., Bülow, L., Hehl, R. (2004) AthaMap: an online resource for in silico transcription factor binding sites in the Arabidopsis thaliana genome, Nucleic Acids Res 32: D368-72

Steffens, N.O., Galuschka, C., Schindler, M., Bülow, L., Hehl, R. (2005) AthaMap web tools for database-assisted identification of combinatorial cis-regulatory elements and the display of highly conserved transcription factor binding sites in Arabidopsis thaliana, Nucleic Acids Res. 33: W397-402

Stevens, R., Mariconti, L., Rossignol, P., Perennes, C., Cella, R., Bergounioux, C. (2002) Two E2F sites in the *Arabidopsis* MCM3 promoter have different roles in cell cycle activation and meristematic expression, J. Biol. Chem. 277:32978-32984

Strasburger, E. (2002) Lehrbuch der Botanik, 35. Auflage, Spektrum Akademischer Verlag Heidelberg Berlin

Sugimoto, K., Takeda, S., Hirochika, H. (2003) Transcriptional activation mediated by binding of a plant GATA-type zinc finger protein AGP1 to the AG-motif (AGATCCAA) of the wound-inducible Myb gene NtMyb2, Plant J. 36: 550-64

Tang, W., Perry, S.E. (2003) Binding site selection for the plant MADS domain protein AGL15: an *in vitro* and *in vivo* study, J. Biol. Chem. 278: 28154-9

Teakle, G.R., Manfield, I.W., Graham, J.F., Gilmartin, P.M. (2002) *Arabidopsis thaliana* GATA factors: organisation, expression and DNA-binding characteristics, Plant Mol. Biol. 50: 43-57

Ulker, B., Somssich, I.E. (2004) WRKY transcription factors: from DNA binding towards biological function, Curr. Opin. Plant Biol. 7: 491-8

Vallelian-Bindschedler, L., Mösinger, E., Métraux, J.P., Schweizer, P. (1998) Structure, expression and localization of a germin-like protein in barley (Hordeum vulgare L.) that is insolubilized in stressed leaves, Plant Mol. Biol. 37:297-308

Wicker, T., Keller, B. (2007) Genome-wide comparative analysis of *copia* retrotransposons in Triticeae, rice, and *Arabidopsis* reveals conserved ancient evolutionary lineages and distinct dynamics of individual copia families, Genome Res. 17:1072-1081

Wilhelm, M. and Wilhelm, F.X. (2001). Reverse transcription of retroviruses and LTR retrotransposons, Cell. Mol. Life Sci**.** 58**:** 1246–1262

Wilusz, C.J., Wilusz, J. (2004) Bringing the role of mRNA decay in the control of gene expression into focus, Trends Genet. 20: 491-7

Yanagisawa, S. (2002), The Dof family of plant transcription factors, Trends Plant Sci. 7: 555-60

Yanagisawa, S. (2004) Dof domain proteins: plant-specific transcription factors associated with diverse phenomena unique to plants, Plant Cell Physiol. 45: 386-91

Yanagisawa, S., Schmidt, R.J. (1999) Diversity and similarity among recognition sequences of Dof transcription factors, Plant J. 17: 209-14

Zimmermann, G., Bäumlein, H., Mock, H.P., Himmelbach, A., Schweizer, P. (2006) The multigene family encoding germin-like proteins of barley. Regulation and function in Basal host resistance, Plant Physiol. 142:181-92

Zimmermann, P., Hennig, L., Gruissem, W. (2005) Gene-expression analysis and network discovery using Genevestigator, Trends Plant Sci. 10: 407-9

Zimmermann, P., Hirsch-Hoffmann, M., Hennig, L., Gruissem, W. (2004) GENEVESTIGATOR. *Arabidopsis* microarray database and analysis toolbox, Plant Physiol. 136:2621-32

7. Anhang

Tabelle 31: Auflistung der pflanzlichen Alignment Matrizen sowie der für die Analysen eliminierten Redundanzen (von zwei putativen Transkriptionsfaktor-Bindungsstellen an einer genomischen Position wurde nur eine verwendet). Redundanzen beziehen sich auf Bindungsstellen, die sowohl auf dem annotierten (+) wie auf dem nicht annotierten (-) DNA-Strang befinden und dieselbe genomische Position beschreiben. Die Zahl bezieht sich auf den Abstand dieser beiden Bindungsstellen voneinander. Von diesen beiden Bindungsstellen wurde jeweils nur eine für die Analysen eingesetzt.

Transkriptionsfaktor	Alignment Matrize	Redundanz
ABF1	A| 20 0 20 0 0 0 0 0 1 C| 0 20 0 20 0 0 0 0 19 G| 0 0 0 0 20 0 20 17 0 T| 0 0 0 0 0 20 0 3 0	-+1
ABI4(1)	A| 0 12 0 0 1 0 1 1 2 4 C| 20 0 21 21 0 21 20 14 13 7 G| 0 9 0 0 20 0 0 5 4 6 T| 1 0 0 0 0 0 0 1 2 4	-
ABI4(2)	A| 0 0 0 0 0 0 4 2 1 0 C| 31 0 0 3 0 28 10 10 18 20 G| 0 31 31 0 31 3 6 8 3 2 T| 0 0 0 28 0 0 11 11 9 9	-
AG_M00151	A | 21 9 10 29 0 0 31 47 52 25 17 19 7 2 22 45 40 15 C | 20 3 0 8 66 65 3 2 0 0 15 8 0 0 17 4 6 10 G | 6 3 1 8 0 0 6 0 1 1 11 20 57 54 5 5 9 16 T | 19 51 55 21 0 1 26 17 13 40 23 19 2 10 22 12 11 25	0
AGL15	A| 5 4 16 0 1 17 10 17 13 19 8 19 0 9 26 19 C| 4 2 2 32 15 4 3 0 0 1 1 0 0 5 1 3 G| 3 1 4 0 0 1 2 1 0 2 4 12 31 4 1 4 T| 20 25 10 0 16 10 17 14 19 10 19 1 1 14 4 6	0
AGL3	A| 22 16 25 27 0 0 82 40 56 35 65 25 64 0 33 52 45 21 C| 29 9 8 16 92 79 1 4 0 0 1 4 0 0 14 5 23 24 G| 34 29 4 13 0 0 2 3 1 0 4 3 28 92 15	0

	7 13 26 T| 10 41 58 39 3 16 10 48 38 60 25 63 3 3 33 31 14 24	
AGP1	A| 5 2 19 0 19 0 0 0 14 4 C| 5 10 0 0 0 0 19 4 0 5 G| 1 3 0 19 0 0 0 0 3 5 T| 8 4 0 0 0 19 0 15 2 5	0
ALFIN1	A| 4 6 3 0 0 0 0 2 1 2 C| 2 2 2 0 0 0 1 1 1 5 G| 5 5 6 13 0 13 8 7 10 4 T| 1 0 2 0 13 0 4 3 1 1	-
ANT	A | 8 0 31 0 21 8 15 0 4 0 0 9 34 7 3 2 C | 0 32 0 34 0 6 1 3 25 34 33 11 0 1 1 10 G | 24 0 2 0 13 15 1 0 0 0 0 13 0 16 29 3 T | 2 2 1 0 0 5 17 31 5 0 1 1 0 10 1 19	-
ATHB1	A| 3 22 25 0 0 25 0 0 3 C| 13 1 0 0 5 0 0 0 1 G| 4 0 0 0 0 0 0 0 12 T| 3 1 0 25 20 0 25 25 0	-
ATHB5	A| 0 4 2 2 0 32 39 0 2 39 0 0 0 C| 9 9 12 17 32 6 0 0 8 0 0 0 0 G| 12 17 8 10 4 0 0 0 7 0 0 0 38 T| 9 4 14 7 3 1 0 39 22 0 39 39 1	+-4
ATHB9_M00417	A | 4 5 5 3 0 0 25 26 0 0 26 0 0 17 0 5 2 0 0 C | 2 3 4 8 1 6 0 0 0 0 0 0 0 0 25 1 6 3 5 G | 1 3 2 2 25 0 1 0 0 26 0 0 3 9 1 4 0 4 3 T | 2 3 5 5 0 20 0 0 26 0 0 26 23 0 0 7 4 3 0	0
AtMYB15	A| 4 27 64 4 0 0 0 87 0 0 0 18 4 C| 73 18 18 64 0 0 0 4 0 0 0 0 0 G| 9 51 18 18 100 100 0 0 100 100 0 78 87 T| 14 4 0 14 0 0 100 9 0 0 100 4 9	-
AtMYB77	A| 33 57 27 6 9 68 0 50 0 0 0 68 3 C| 30 17 9 30 6 3 100 0 0 0 0 3 55 G| 20 17 37 20 82 23 0 50 100 0 0 26 30 T| 17 9 27 44 3 6 0 0 0 100 100 3 12	-
AtMYB84	A| 22 0 18 14 0 0 0 78 0 0 0 0 4 C| 18 14 4 9 0 0 0 4 0 0 0 0 32 G| 42 86 78 55 100 100 0 4 100 100 0 96 60	-

	T| 18 0 0 22 0 0 100 14 0 0 100 4 4	
bZIP910(1)	A| 0 14 0 0 18 0 0 0 0 0 0 10 C| 0 0 0 0 0 18 0 0 0 0 18 8 G| 18 4 0 18 0 0 18 0 18 17 0 0 T| 0 0 18 0 0 0 0 18 0 1 0 0	0
bZIP910(2)	A| 0 1 6 2 0 1 0 0 17 0 0 0 C| 0 4 1 1 1 15 0 0 0 17 0 0 G| 17 10 9 2 16 1 0 17 0 0 17 0 T| 0 2 1 12 0 0 17 0 0 0 0 17	+-8
bZIP911(1)	A| 1 10 0 0 22 0 0 0 0 0 4 1 C| 0 0 0 0 0 22 0 0 0 0 18 21 G| 21 12 0 22 0 0 22 0 22 22 0 0 T| 0 0 22 0 0 0 0 22 0 0 0 0	0
bZIP911(2)	A| 0 6 0 0 10 0 0 0 0 0 8 0 C| 0 0 0 0 0 10 0 0 0 0 1 10 G| 10 4 0 10 0 0 10 0 10 0 1 0 T| 0 0 10 0 0 0 0 10 0 10 0 0	0
CBF	A| 41 45 35 39 41 0 131 131 0 37 42 38 52 C| 25 22 21 24 45 131 0 0 0 26 26 33 22 G| 26 26 35 28 20 0 0 0 0 26 24 20 20 T| 39 38 39 41 26 0 0 0 131 42 39 41 37	-
CDC5_M00361	A | 28 17 0 2 6 91 2 2 4 2 18 C | 24 26 92 2 85 2 0 93 4 55 23 G | 32 55 4 0 0 2 93 5 91 27 45 T | 15 2 4 96 8 4 4 0 0 16 14	-
DOF2_M00353	A | 7 7 7 7 21 21 21 0 3 7 7 C | 7 6 3 2 0 0 0 0 12 4 6 G | 2 4 4 2 0 0 0 21 2 6 5 T | 5 4 7 10 0 0 0 0 4 4 3	-
GAMYB	A | 0 25 25 0 3 5 10 5 C | 16 0 0 25 13 3 10 18 G | 0 0 0 0 7 17 1 2 T | 9 0 0 0 2 0 4 0	-
GT1	A| 9 7 10 5 4 0 1 11 17 14 11 9 C| 1 1 2 3 2 0 0 0 0 1 0 2 G| 7 4 4 3 9 18 0 1 0 0 0 0 T| 1 6 2 7 3 0 17 6 1 3 7 7	-
HAHB4	A| 1 25 25 0 7 25 0 3 6 C| 0 0 0 0 0 0 0 1 5 G| 0 0 0 0 14 0 0 3 14 T| 24 0 0 25 4 0 25 18 0	-
HVH21	A | 3 14 5 4 0 1 37 1 15 4 9 7 C | 12 11 12 14 3 1 0 35 8 8 6 4 G | 11 9 7 16 1 35 0 0 13 18 13	-

	11 T | 11 3 13 3 33 0 0 1 1 7 9 15	
ID1	A| 0 0 0 0 0 0 0 0 4 5 3 C| 3 0 1 0 0 14 5 6 2 1 1 G| 0 0 0 14 1 0 8 0 0 0 0 T| 12 15 14 1 14 1 2 9 9 9 11	-
MYB.PH3_M00218	A | 22 21 27 40 33 33 0 3 0 0 0 46 1 C | 3 1 1 0 0 0 46 15 0 0 0 0 13 G | 4 7 7 2 8 8 0 26 46 0 0 0 2 T | 11 12 6 3 4 5 0 2 0 46 46 0 7	-
MYB.PH3_M00219	A | 12 12 11 25 0 0 0 23 0 0 0 23 1 C | 1 2 2 0 0 0 0 0 0 0 0 0 5 G | 6 7 8 0 27 3 0 0 27 0 0 4 4 T | 6 4 4 1 0 24 27 4 0 27 27 0 1	-
O2	A| 2 3 0 11 0 0 0 0 5 2 C| 3 4 8 0 11 0 0 4 0 4 G| 4 0 3 0 0 11 0 7 3 3 T| 2 4 0 0 0 0 11 0 3 2	0
P	A| 25 2 0 17 36 0 1 11 6 C| 3 28 31 0 0 36 34 13 8 G| 4 3 0 0 0 0 0 4 15 T| 4 3 5 19 0 0 1 8 7	-
PCF2	A | 15 16 3 3 17 2 0 0 64 1 C | 14 14 1 6 19 71 73 75 1 66 G | 11 30 66 65 25 2 3 0 8 2 T | 24 14 6 2 15 1 0 1 3 7	-+1
PCF5	A | 6 4 0 1 5 0 0 4 17 15 C | 3 20 0 0 22 83 83 74 24 28 G | 63 5 82 82 13 0 0 3 25 12 T | 3 52 1 0 43 0 0 2 15 26	0; -+1; -+2
PIF3_M00434	A| 2 5 11 6 5 11 0 27 0 0 0 0 11 9 3 11 4 4 C| 5 1 2 0 3 11 27 0 27 0 0 0 2 12 9 6 19 2 G| 15 10 12 17 17 5 0 0 0 27 0 27 13 6 13 0 4 9 T| 5 11 2 4 2 0 0 0 0 0 27 0 1 0 2	0

	10 0 12	
PIF3(2)	A| 15 9 10 8 1 0 32 0 0 0 0 0 9 15 11 7 4 C| 2 4 2 11 26 32 0 32 0 0 0 0 10 15 13 8 11 G| 9 10 13 13 5 0 0 0 32 0 32 32 11 0 8 11 6 T| 6 9 7 0 0 0 0 0 0 32 0 0 2 2 0 6 11	-+1
RAV1_M00343	A| 14 16 7 0 47 62 0 69 9 41 34 27 C| 17 11 14 62 12 4 69 0 7 4 6 9 G| 12 10 35 0 9 3 0 0 32 4 10 17 T| 6 14 1 0 0 0 0 0 21 20 19 16	-
RAV1_M00344	A| 22 13 8 40 1 0 0 0 24 12 12 8 C| 12 15 49 6 56 65 0 0 7 15 14 20 G| 10 8 1 8 3 0 0 58 19 18 16 5 T| 18 28 7 11 5 0 65 0 5 9 10 13	-
TBP	A | 48 27 5 162 0 171 106 166 65 125 22 51 C | 46 108 2 0 7 0 0 0 2 14 72 72 G | 29 9 0 0 0 0 0 3 0 17 48 27 T | 48 27 164 9 164 0 65 2 104 15 31 19	0
TEIL	A | 70 1 0 31 86 1 5 3 C | 6 21 1 12 0 49 61 8 G | 6 0 86 3 1 2 13 4 T | 5 65 0 41 0 35 8 72	-
TGA1	A | 10 6 7 12 0 0 31 0 0 7 6 18 C | 4 6 8 8 0 0 0 31 0 3 3 0 G | 6 6 13 11 0 31 0 0 31 2 6 7 T | 11 13 3 0 31 0 0 0 0 19 16 6	+-4, -+4
TGA1a	A| 1 3 0 11 0 0 0 0 5 1 C| 5 3 8 0 11 0 0 4 1 4 G| 4 1 3 0 0 11 0 7 2 5 T| 1 4 0 0 0 0 11 0 3 1	0
ZAP1	A | 0 0 0 45 0 0 1 33 2 C | 0 0 0 0 45 44 3 6 4 G | 0 0 45 0 0 0 39 3 35 T | 45 45 0 0 0 1 2 3 4	-

Tabelle 32: Auflistung der nichtpflanzlichen Alignment Matrizen, extrahiert aus der öffentlich zugänglichen Version von TRANSFAC©, sowie sowie der für die Analysen eliminierten Redundanzen (von zwei putativen Transkriptionsfaktor-Bindungsstellen an einer genomischen Position wurde nur eine verwendet). Redundanzen beziehen sich auf Bindungsstellen, die sowohl auf dem annotierten (+) wie auf dem nicht annotierten (-) DNA-Strang befinden und dieselbe genomische Position beschreiben. Die Zahl bezieht sich auf den Abstand dieser beiden Bindungsstellen voneinander. Von diesen beiden Bindungsstellen wurde jeweils nur eine für die Analysen eingesetzt.

Transkriptionsfaktor	**Alignment Matrize**	**Redundanz**
abaA	A| 3 5 2 0 0 1 7 0 22 0 0 0 0 7 5 4 3 6 1 C| 10 9 6 6 10 8 2 22 0 0 0 22 15 5 4 7 6 9 7 G| 2 6 5 6 2 1 7 0 0 0 0 0 0 2 7 3 8 4 5 T| 7 2 9 10 10 12 6 0 0 22 22 0 7 8 6 8 5 3 9	-
AP-2alpha	A| 4 3 0 0 0 2 2 4 1 2 3 1 C| 6 0 11 13 12 5 8 5 1 2 9 3 G| 3 6 0 0 0 6 3 3 10 9 1 8 T| 0 4 2 0 1 0 0 1 1 0 0 1	0
AREB6_M00414	A| 3 1 5 0 12 0 0 0 0 0 3 3 C| 5 4 0 12 0 11 11 0 0 1 1 6 G| 4 2 7 0 0 1 0 12 5 5 2 T| 0 5 0 0 0 1 0 12 0 6 3 1	0
AREB6_M00415	A| 3 0 0 0 0 0 0 4 3 C| 6 3 0 0 0 0 7 2 2 G| 2 4 12 0 0 0 5 2 4 T| 1 5 0 12 12 12 0 4 3	-
Brachyury	A| 15 5 15 5 1 2 30 0 38 2 2 1 40 0 0 0 0 0 1 40 31 32 5 1 C| 17 3 14 12 2 19 0 40 0 34 28 1 0 0 0 0 0 2 7 0 5 0 5 5 G| 3 15 0 16 0 19 7 0 2 2 8 0 0 40 40 0 40 0 28 0 0 3 1 8 T| 5 17 11 7 37 0 3 0 0 2 2 38 0 0 0 40 0 38 4 0 4 5 29 26	0
c-Ets-1	A| 7 1 5 0 0 15 8 4 1 5 C| 2 14 9 0 0 0 2 1 6 4 G| 3 0 0 15 15 0 0 10 0 4 T| 3 0 1 0 0 0 5 0 8 2	-
c-Ets-154	A| 9 9 19 10 13 0 0 40 23 15 10 9 11 C| 11 11 4 20 18 0 0 0 0 5 8 5 9 G| 9 6 4 6 2 40 40 0 0 15 1 7 7 T| 4 9 9 2 5 0 0 0 17 4 20 15 8	-

COUP	A| 0 1 12 6 0 0 0 1 2 6 6 1 3 0 C| 0 0 0 7 13 3 2 0 0 4 5 10 6 3 G| 2 12 1 0 0 0 0 0 11 3 1 1 0 3 T| 11 0 0 0 0 10 11 12 0 0 1 1 4 7	-
CP2	A| 0 0 2 2 2 5 2 0 2 6 0 C| 1 5 2 4 0 1 3 6 4 0 0 G| 5 1 0 0 2 0 1 0 0 0 6 T| 0 0 2 0 2 0 0 0 0 0 0	-
DI	A| 3 0 5 0 12 11 12 11 3 0 4 C| 4 3 2 0 0 0 0 0 4 8 5 G| 0 9 5 12 0 1 0 0 1 1 3 T| 5 0 0 0 0 0 0 1 4 3 0	-
E2F	A| 1 2 1 0 0 0 0 0 5 5 5 5 1 0 2 C| 0 1 2 1 5 0 4 0 0 0 0 0 2 0 0 G| 0 0 2 4 0 5 1 5 0 0 0 0 0 2 3 T| 4 2 0 0 0 0 0 0 0 0 0 0 2 3 0	-+5
E74A	A| 4 3 11 11 2 2 5 0 0 17 17 5 1 3 5 2 2 C| 5 3 1 0 7 12 12 0 0 0 0 1 2 2 3 1 3 G| 2 6 2 2 2 2 0 17 17 0 0 10 1 7 4 6 6 T| 3 2 1 3 6 1 0 0 0 0 0 1 11 3 3 5 1	-
FOXA4a	A| 1 4 2 2 0 6 6 6 0 6 3 2 2 0 C| 3 0 1 0 0 0 0 0 2 0 1 0 0 3 G| 0 1 0 4 0 0 0 0 0 0 0 1 3 1 T| 2 1 3 0 6 0 0 0 4 0 2 3 1 2	-
GAL4	A| 1 5 3 1 0 0 4 1 2 7 1 4 1 0 1 1 2 0 0 0 8 7 2 C| 5 2 2 10 0 1 3 3 4 0 8 1 3 2 6 5 1 10 11 0 0 0 6 G| 3 1 1 0 10 10 3 4 4 2 2 0 5 1 2 4 1 1 0 11 0 4 3 T| 2 3 5 0 1 0 1 3 1 2 0 6 2 8 2 1 7 0 0 0 3 0 0	0
GCM	A| 36 33 68 0 0 10 0 2 100 2 43 24 26 C| 28 15 10 98 100 88 2 92 0 9 26 24 16 G| 28 26 32 0 0 0 98 0 0 2 17 10 21 T| 8 26 0 2 0 2 0 6 0 87 14 43 37	-
HAP3	A| 7 1 1 4 6 0 0 14 14 0 4 13 3 7 4 5 C| 2 6 9 5 0 14 14 0 0 2 6 0 5 5 3 3 G| 2 0 3 5 4 0 0 0 0 0 2 1 5 2 5 4 T| 3 7 1 0 4 0 0 0 0 12 2 0 1 0 2 2	-
LAC9	A| 14 0 15 13 0 0 0 72 57 15 56 0 13 0 0 41 27 15 0 0 0 30 C| 27 0 15 0 100 0 0 14 14 28 0 70 0 13 41 14 43 0 100 100 14 15 G| 30 0 0 30 0 100 100 14 0 15 0 0 0 45 14 30 30 0 0 0 85 28	+-3

	T| 29 100 0 57 0 0 0 0 28 42 44 30 72 42 45 15 0 85 0 0 0 27	
NF-kappaB1	A| 4 2 0 0 1 13 6 1 0 1 1 0 2 3 C| 1 0 0 0 0 0 6 1 3 3 12 12 4 4 G| 4 8 11 13 12 0 0 2 0 0 0 0 3 3 T| 4 3 2 0 0 0 1 9 10 9 0 1 4 3	0; +-1
Pax-2	A| 7 5 7 8 8 5 1 21 6 3 3 14 4 4 18 3 1 3 3 C| 10 10 8 10 1 2 20 3 13 3 13 1 5 3 5 8 5 3 3 G| 4 7 2 6 19 0 3 3 3 20 7 11 3 14 2 5 7 5 2 T| 6 9 15 8 4 25 8 5 10 6 7 4 18 8 4 9 6 2 2	+-1
Pax-3	A| 0 0 0 4 0 25 0 11 0 4 1 9 13 C| 0 26 0 0 25 1 24 0 21 5 8 10 8 G| 0 0 26 0 0 0 0 15 4 1 0 0 3 T| 26 0 0 22 1 0 2 0 1 16 17 7 2	-
Pax-4a	A| 4 1 6 11 3 0 0 36 11 0 3 3 0 7 8 6 4 10 8 2 6 C| 7 4 16 15 3 4 29 1 5 4 30 3 7 5 11 4 14 5 6 13 9 G| 8 26 10 16 25 3 14 6 6 34 7 33 7 23 5 12 11 13 8 4 6 T| 7 4 10 0 12 40 0 0 21 5 3 4 27 3 8 8 1 3 7 8 4	-
Pax-5_M00143	A| 0 1 1 3 2 3 1 0 0 7 2 0 0 3 2 0 2 3 1 4 1 2 1 2 1 2 1 0 C| 2 5 2 1 1 0 3 1 7 0 2 2 1 1 0 0 3 0 0 1 1 5 3 0 4 3 3 3 G| 2 0 3 1 3 4 2 4 0 0 2 2 6 1 1 7 1 4 2 2 5 0 3 4 2 0 2 2 T| 3 1 1 2 1 0 1 2 0 0 1 3 0 2 4 0 1 0 4 0 0 0 0 1 0 2 1 2	-
Pax-5_M00144	A| 2 3 2 1 2 1 3 1 2 0 0 0 1 3 4 0 0 0 0 2 4 0 1 3 0 2 1 2 C| 1 0 2 2 0 0 0 1 0 2 3 1 1 0 0 0 5 0 0 0 0 5 2 0 2 1 2 2 G| 2 2 0 2 1 4 1 1 1 1 1 0 2 2 1 5 0 5 2 3 0 0 2 2 0 1 2 0 T| 0 0 1 0 2 0 1 2 2 2 1 4 1 0 0 0 0 0 3 0 1 0 0 0 3 1 0 1	-
Pax-6	A| 15 21 10 8 3 2 3 40 3 1 1 23 1 2 40 14 2 1 13 14 4 C| 7 9 9 14 2 0 29 5 39 0 36 2 4 13 1 11 4 0 17 8 12	-+1

	G| 6 3 10 9 4 1 1 1 0 44 7 1 0 26 6 15 3 20 9 4 3 T| 10 10 18 16 38 44 14 1 5 2 2 21 42 3 0 7 37 25 4 6 9	
POU2F1_M00136	A| 14 10 3 33 36 2 44 0 0 0 33 12 11 10 15 C| 6 5 5 5 2 1 0 1 1 43 3 10 7 11 7 G| 8 8 27 4 2 0 0 0 23 0 5 6 11 6 6 T| 4 13 6 2 4 41 0 43 20 1 3 13 11 13 12	-+1
POU2F1_M00162	A| 1 3 2 4 3 0 0 3 1 6 0 2 0 2 C| 4 0 1 1 1 0 0 0 3 0 0 0 3 1 G| 1 1 1 0 0 0 3 1 2 0 0 3 0 1 T| 0 2 2 1 2 6 3 2 0 0 6 1 3 2	+-1
PPAR-alpha	A| 1 4 4 5 4 1 0 6 0 0 1 1 7 5 7 0 0 0 0 6 C| 5 1 0 1 0 5 2 0 0 0 2 6 0 0 0 0 0 0 6 0 G| 1 0 3 1 1 1 0 1 7 7 1 0 0 0 0 7 5 0 1 0 T| 0 2 0 0 2 0 5 0 0 0 3 0 0 2 0 0 2 7 0 1	-

Tabelle 33: Zufallsmatrizen für die pflanzlichen Matrizen, erzeugt mit SeqToMatrix. Die Länge der Ursprungsmatrize, die Anzahl der in die Matrize eingehenden Sequenzen sowie der geradzahlige AT-Gehalt wurden bei der Erstellung berücksichtigt

Transkriptionsfaktor	**Name der Matrize**	**Alignment Matrize**
ABF1	randomABF1	A| 1 4 0 7 9 4 2 3 5 C| 12 4 3 4 4 4 2 9 4 G| 4 10 8 4 4 9 13 4 7 T| 3 2 9 5 3 3 3 4 4
ABI4(1)	randomABI4(1)	A| 1 0 1 1 2 1 2 1 4 1 C| 6 8 10 7 5 8 13 9 8 5 G| 12 12 9 12 14 9 5 9 7 11 T| 2 1 1 1 0 3 1 2 2 4
ABI4(2)	randomABI4(2)	A| 4 1 6 3 3 4 1 4 2 5 C| 14 13 6 12 13 10 11 10 8 10 G| 7 15 14 13 13 15 15 14 15 10 T| 6 2 5 3 2 2 4 3 6 6
AG_M00151	randomAG	A| 18 22 28 25 24 15 25 19 19 22 16 20 20 17 19 19 21 22 C| 14 11 8 13 12 11 8 16 14 12 13 15 10 18 8 13 14 10 G| 13 11 10 18 10 14 11 10 12 12 13 8 11 9 10 10 13 8 T| 21 22 20 10 20 26 22 21 21 20 24 23 25 22 29 24 18 26
AGL15	randomAGL15	A| 9 12 9 7 11 14 11 11 10 15 15 13 10 11 11 7 C| 2 5 4 4 5 3 1 8 3 0 0 3 5

		5 4 4 G| 8 5 8 4 4 7 7 4 9 4 4 6 4 3 6 5 T| 13 10 11 17 12 8 13 9 10 13 13 10 13 13 11 16
AGL3	randomAGL3	A| 31 31 35 31 33 26 31 27 28 26 33 25 31 27 32 25 30 28 C| 20 15 14 17 17 17 17 20 13 14 24 19 15 16 13 23 20 19 G| 17 13 12 18 15 16 17 22 20 15 13 11 14 12 10 12 20 15 T| 27 36 34 29 30 36 30 26 34 40 25 40 35 40 40 35 25 33
AGP1	randomAGP1	A| 8 4 2 5 3 6 3 9 7 3 C| 2 5 4 2 7 5 3 2 2 3 G| 3 5 4 5 2 2 8 5 4 4 T| 6 5 9 7 7 6 5 3 6 9
ALFIN1	randomALFIN1	A| 2 4 0 3 1 4 2 1 2 3 C| 7 5 6 3 5 4 2 4 6 2 G| 2 1 7 3 3 5 4 7 3 8 T| 2 3 0 4 4 0 5 1 2 0
ANT	randomANT	A| 9 7 7 9 10 5 11 9 9 6 9 2 11 9 4 4 C| 12 13 9 6 7 12 7 10 9 12 11 17 8 8 9 8 G| 7 7 10 12 9 11 11 7 8 11 8 7 9 10 10 8 T| 6 7 8 7 8 6 5 8 8 5 6 8 6 7 11 14
ATHB1	randomATHB1	A| 14 11 11 10 9 13 12 9 14 C| 4 3 2 0 2 1 0 1 2 G| 1 1 4 0 2 2 4 3 0 T| 6 10 8 15 12 9 9 12 9
ATHB5	randomATHB5	A| 8 16 7 14 16 14 12 20 12 15 10 14 17 C| 12 8 6 7 6 8 11 7 8 3 6 6 5 G| 6 3 14 5 5 7 9 4 4 9 9 9 7 T| 13 12 12 13 12 10 7 8 15 12 14 10 10
ATHB9_M00417	randomATHB9	A| 6 9 10 9 9 7 10 8 6 5 11 7 10 7 7 10 8 12 7 C| 5 3 3 3 3 7 3 8 6 1 3 5 7 6 5 6 7 1 4 G| 6 6 3 6 4 2 8 4 9 15 7 6 3 5 8 2 5 6 4 T| 9 8 10 8 10 10 5 6 5 5 5 8 6 8 6 8 6 7 11
AtMYB15	randomAtMYB15	A| 19 15 21 21 17 25 17 20 16 21 17 20 18

		C| 30 33 20 38 36 28 31 23 38 32 34 31 40 G| 33 35 27 27 33 30 32 35 30 26 24 28 27 T| 18 17 32 14 14 17 20 22 16 21 25 21 15
AtMYB77	randomAtMYB77	A| 26 23 27 33 28 24 26 28 31 24 22 28 25 C| 19 22 25 19 24 31 18 24 21 29 37 22 30 G| 26 29 30 28 29 26 31 30 21 24 18 27 19 T| 29 26 18 20 19 19 25 18 27 23 23 23 26
AtMYB84	randomAtMYB84	A| 17 14 13 11 19 18 21 17 19 14 17 15 7 C| 31 30 43 43 32 27 29 36 31 33 21 26 38 G| 36 42 30 30 38 43 36 34 35 24 42 45 42 T| 16 14 14 16 11 12 14 13 15 29 20 14 13
bZIP910(1)	randombZIP910(1)	A| 2 2 4 3 5 4 5 3 8 4 3 0 C| 4 9 7 3 5 5 5 3 2 4 6 9 G| 6 5 5 7 5 8 5 9 7 5 7 5 T| 6 2 2 5 3 1 3 3 1 5 2 4
bZIP910(2)	randombZIP910(2)	A| 4 1 3 3 4 2 3 4 2 1 3 4 C| 7 4 6 7 8 5 4 6 6 8 2 4 G| 3 8 5 3 3 6 4 4 4 5 8 5 T| 3 4 3 4 2 4 6 3 5 3 4 4
bZIP911(1)	randombZIP911(1)	A| 4 3 2 2 1 5 3 5 5 4 3 3 C| 9 12 10 14 9 9 6 10 6 6 7 4 G| 5 4 9 5 8 6 8 6 9 9 9 9 T| 4 3 1 1 4 2 5 1 2 3 3 6
bZIP911(2)	randombZIP911(2)	A| 2 1 3 2 5 3 3 1 2 5 3 3 C| 2 4 3 3 2 3 2 6 3 0 4 2 G| 5 2 3 2 3 2 1 2 4 3 2 2 T| 1 3 1 3 0 2 4 1 1 2 1 3
CBF	randomCBF	A| 44 35 44 45 38 45 41 33 35 45 43 45 45 C| 24 23 25 20 24 18 18 27 20 22 20 23 21 G| 28 25 21 18 31 29 22 25 26 27 26 22 23 T| 35 48 41 48 38 39 50 46 50 37 42 41 42
CDC5_M00361	randomCDC5	A| 20 17 17 17 11 17 10 14 17 17 10 C| 31 30 29 32 34 31 45 36 28 32 31 G| 31 41 37 35 30 36 33 34 41 34 36 T| 17 11 16 15 24 15 11 15 13 16 22
DOF2_M00353	randomDOF2	A| 8 6 7 6 7 8 11 5 7 9 8 C| 1 3 1 2 6 1 3 4 5 2 3 G| 6 5 6 5 2 2 3 6 3 3 3 T| 6 7 7 8 6 10 4 6 6 7 7
GAMYB	randomGAMYB	A| 6 5 6 6 5 7 4 3 C| 3 4 8 4 12 4 7 10 G| 8 6 6 11 5 7 10 8 T| 8 10 5 4 3 7 4 4
GT1	randomGT1	A| 5 7 6 6 8 5 6 6 8 8 8 6 C| 2 2 4 3 2 1 2 1 5 3 2 3 G| 4 2 0 3 1 3 2 3 0 3 2 4

		T| 7 7 8 6 7 9 8 8 5 4 6 5
HAHB4	randomHAHB4	A| 7 13 8 11 9 6 14 15 8 C| 2 1 3 1 2 2 2 0 3 G| 1 0 1 1 1 4 0 2 2 T| 15 11 13 12 13 13 9 8 12
HVH21	randomHVH21	A| 8 13 8 7 6 4 7 1 9 12 6 9 C| 9 7 9 9 10 5 7 12 10 12 13 13 G| 10 8 10 14 12 20 10 9 7 8 11 10 T| 10 9 10 7 9 8 13 15 11 5 7 5
ID1	randomID1	A| 7 5 7 2 3 4 1 5 7 6 5 C| 3 2 0 2 3 2 2 2 1 6 1 G| 1 2 3 5 5 3 4 0 2 2 3 T| 4 6 5 6 4 6 8 8 5 1 6
MYB.PH3_M00218	randomMYB.PH3(1)	A| 16 10 15 18 20 13 13 16 15 14 13 10 16 C| 5 9 5 8 5 12 13 13 8 5 11 8 5 G| 12 7 6 5 7 8 3 6 10 9 5 9 12 T| 13 20 20 15 14 13 17 11 13 18 17 19 13
MYB.PH3_M00219	randomMYB.PH3(2)	A| 8 6 8 11 8 14 7 7 11 11 7 6 7 C| 2 4 5 5 4 2 7 4 2 3 6 5 5 G| 4 7 5 2 3 5 2 6 5 5 1 6 3 T| 13 10 9 9 12 6 11 10 9 8 13 10 12
O2	randomO2	A| 2 2 3 4 2 4 2 2 4 2 C| 4 1 1 4 6 3 4 5 4 3 G| 4 6 6 2 2 2 2 3 2 2 T| 1 2 1 1 1 2 3 1 1 4
P	randomP	A| 10 11 9 6 9 9 10 8 11 C| 10 9 11 12 7 7 10 14 12 G| 10 8 8 10 10 17 7 6 8 T| 6 8 8 8 10 3 9 8 5
PCF2	randomPCF2	A| 7 10 12 8 7 9 6 5 8 12 C| 32 28 29 29 28 29 29 29 33 29 G| 27 28 25 29 32 30 28 33 26 25 T| 10 10 10 10 9 8 13 9 9 10
PCF5	randomPCF5	A| 10 9 12 8 10 11 9 12 14 6 C| 34 34 25 34 24 28 30 21 30 31 G| 30 29 39 27 34 32 31 37 26 36 T| 9 11 7 14 15 12 13 13 13 10
PIF3_M00434	randomPIF3(1)	A| 6 5 6 3 6 4 4 5 5 2 4 3 6 8 4 6 5 6 C| 5 11 6 7 7 10 7 10 9 5 9 9 9 8 9 10 6 8 G| 11 9 8 11 10 6 11 7 11 13 10 9 9 5 9 8 8 7 T| 5 2 7 6 4 7 5 5 2 7 4 6 3 6 5 3 8 6
PIF3(2)	randomPIF3(2)	A| 8 4 6 4 4 5 5 5 4 5 9 9 7

		3 1 8 5 C| 9 10 13 13 14 7 15 13 10 9 8 10 9 9 13 7 9 G| 10 8 12 8 8 10 6 10 12 12 9 10 9 13 11 12 12 T| 5 10 1 7 6 10 6 4 6 6 6 3 7 7 7 5 6
RAV1_M00343	randomRAV1(1)	A| 15 19 20 14 20 22 13 20 20 17 17 23 C| 9 14 18 19 20 19 17 14 9 14 16 14 G| 21 15 14 24 12 9 23 14 12 18 12 17 T| 24 21 17 12 17 19 16 21 28 20 24 15
RAV1_M00344	randomRAV1(2)	A| 16 19 12 14 13 16 14 12 13 20 15 21 C| 14 17 16 17 26 14 20 24 15 18 17 19 G| 21 17 22 16 12 26 15 11 25 9 19 16 T| 14 12 15 18 14 9 16 18 12 18 14 9
TBP	randomTBP	A| 66 60 66 69 74 60 69 65 70 70 62 74 C| 14 22 15 20 18 25 25 16 16 17 21 15 G| 15 15 20 19 17 14 22 21 18 13 16 20 T| 76 74 70 63 62 72 55 69 67 71 72 62
TEIL	randomTEIL	A| 21 27 21 29 24 19 31 31 C| 21 22 17 14 18 16 15 21 G| 19 15 20 21 13 18 12 18 T| 26 23 29 23 32 34 29 17
TGA1	randomTGA1	A| 14 10 7 6 7 11 8 4 6 5 7 6 C| 8 5 9 9 8 7 9 11 9 5 8 17 G| 3 8 8 7 7 9 6 6 10 10 7 4 T| 6 8 7 9 9 4 8 10 6 11 9 4
TGA1a	randomTGA1a	A| 1 1 3 2 2 2 1 7 1 1 C| 3 4 1 0 4 2 5 1 5 4 G| 5 1 3 6 4 5 4 0 3 6 T| 2 5 4 3 1 2 1 3 2 0
ZAP1	randomZAP1	A| 6 8 5 8 12 10 9 12 11 C| 15 18 13 11 13 10 16 10 12 G| 13 12 15 14 12 13 13 12 11 T| 11 7 12 12 8 12 7 11 11

Tabelle 34: Zufallsmatrizen für die nichtpflanzlichen Matrizen, erzeugt mit SeqToMatrix. Die Länge der Ursprungsmatrize, die Anzahl der in die Matrize eingehenden Sequenzen sowie der geradzahlige AT-Gehalt wurden bei der Erstellung berücksichtigt

Transkriptionsfaktor	**Name der Matrize**	**Alignment Matrize**
abaA	randomabaA	A| 6 2 5 8 4 6 4 3 5 9 5 4 8 8 6 6 8 4 6 C| 8 9 6 5 5 5 6 12 7 4 6 4 4 4 5 7 3 2 5

		G| 4 7 8 5 8 6 8 6 7 7 6 6 5 6 7 5 7 10 5 T| 4 4 3 4 5 5 4 1 3 2 5 8 5 4 4 4 4 6 6
AP-2alpha	randomAP-2alpha	A| 1 4 2 3 3 0 2 1 1 3 3 3 C| 6 5 5 2 3 5 4 7 6 4 4 5 G| 5 4 4 5 5 6 7 3 4 5 5 5 T| 1 0 2 3 2 2 0 2 2 1 1 0
AREB6_M00414	randomAREB6_M00414	A| 2 2 2 1 3 1 4 2 1 3 5 2 C| 5 3 2 3 2 4 3 4 2 5 3 2 G| 3 4 3 6 3 5 5 5 6 4 2 4 T| 2 3 5 2 4 2 0 1 3 0 2 4
AREB6_M00415	randomAREB6_M00415	A| 3 2 2 5 4 6 2 3 3 C| 2 1 2 0 2 1 3 5 4 G| 4 3 4 2 2 3 2 2 2 T| 3 6 4 5 4 2 5 2 3
Brachyury	randomBrachyury	A| 11 8 9 13 8 14 9 11 9 15 8 13 11 14 12 8 14 8 11 14 14 15 9 17 C| 8 1 9 8 10 10 17 10 10 8 9 11 7 8 9 6 8 10 6 8 10 12 8 4 G| 11 18 7 8 8 7 9 5 12 4 8 9 10 13 9 15 6 9 11 8 11 5 7 7 T| 10 13 15 11 14 9 5 14 9 13 15 7 12 5 10 11 12 13 12 10 5 8 16 12
c-Ets-1	randomc-Ets-1	A| 6 5 3 2 3 5 2 4 2 4 C| 3 7 5 6 2 4 4 4 4 4 G| 4 2 4 3 4 2 6 2 4 4 T| 2 1 3 4 6 4 3 5 5 3
c-Ets-154	randomc-Ets-154	A| 8 13 14 9 12 13 6 9 9 10 11 8 11 C| 9 9 5 8 11 9 7 14 9 9 10 6 9 G| 9 8 9 11 6 6 12 10 15 9 10 8 12 T| 14 10 12 12 11 12 15 7 7 12 9 18 8
COUP	randomCOUP	A| 3 4 1 3 3 4 5 3 5 5 3 6 4 5 C| 5 3 4 2 4 6 2 2 5 3 4 3 2 1 G| 2 4 5 4 4 1 2 2 2 1 4 2 4 4 T| 3 2 3 4 2 2 4 6 1 4 2 2 3 3
CP2	randomCP2	A| 1 0 2 2 0 1 0 0 1 2 1 C| 2 2 0 0 2 1 3 3 1 2 2

		G| 2 2 3 2 2 3 2 1 2 2 2 T| 1 2 1 2 2 1 1 2 2 0 1
DI	randomDI	A| 3 5 4 5 4 4 4 2 2 4 4 C| 3 2 4 1 2 1 4 0 4 1 1 G| 2 0 4 2 4 5 4 6 5 2 1 T| 4 5 0 4 2 2 0 4 1 5 6
E2F	randomE2F	A| 1 3 1 0 2 1 1 1 1 3 2 1 1 1 3 C| 3 1 1 1 2 2 2 1 0 1 2 2 0 2 0 G| 1 0 0 0 1 0 1 3 2 0 0 1 2 1 1 T| 0 1 3 4 0 2 1 0 2 1 1 1 2 1 1
E74A	randomE74A	A| 4 5 4 1 6 1 2 4 4 3 3 3 6 2 4 7 4 C| 4 4 4 5 4 5 3 3 5 5 6 7 3 3 5 2 6 G| 4 3 4 7 5 6 6 5 3 4 4 2 4 1 3 3 5 T| 5 5 5 4 2 5 6 5 5 5 4 5 4 11 5 5 2
FOXA4a	randomFOXA4a	A| 0 3 1 2 2 4 2 3 4 3 2 3 3 2 C| 0 0 0 1 0 0 0 2 1 0 1 0 0 2 G| 2 1 2 2 1 1 1 1 0 1 1 0 0 0 T| 4 2 3 1 3 1 3 0 1 2 2 3 3 2
GAL4	randomGAL4	A| 5 1 5 2 0 3 1 1 3 1 3 3 4 5 2 1 1 3 3 3 4 2 3 C| 2 1 1 4 2 3 5 2 3 4 3 1 1 3 5 6 2 2 3 3 4 4 4 G| 2 5 2 4 7 1 3 6 3 2 3 4 5 3 2 3 3 4 2 3 1 4 3 T| 2 4 3 1 2 4 2 2 2 4 2 3 1 0 2 1 5 2 3 2 2 1 1
GCM	randomGCM	A| 16 22 17 19 23 22 22 23 16 23 20 18 16 C| 28 29 32 29 26 33 25 29 32 19 28 30 32 G| 31 32 25 25 34 21 29 29 30 29 31 30 31 T| 25 17 26 27 17 24 24 19 22 29 21 22 21
HAP3	randomHAP3	A| 5 1 2 2 4 3 3 5 3 2 4 5 5 6 8 6 C| 3 0 4 2 4 5 5 2 5 4 2 1 3 2 3 1 G| 5 8 5 8 4 4 3 4 1 3 4 3 6 0 1 3 T| 1 5 3 2 2 2 3 3 5 5 4 5 0 6 2 4
LAC9	randomLAC9	A| 24 25 22 16 18 19 34 21 24 24 22 23 25 26 19 25 25 21 27 23 22 26 C| 23 26 23 32 30 26 22 27 31 23 25 25 29 23 28 30 25 29 24 32 43 30

		G| 35 20 28 27 36 28 24 37 25 25 30 28 27 29 29 23 25 26 25 27 18 23 T| 18 29 27 25 16 27 20 15 20 28 23 24 19 22 24 22 25 24 24 18 17 21
NF-kappaB1	randomNF-kappaB1	A| 5 0 2 1 4 3 2 6 4 4 2 3 3 3 C| 4 5 4 5 5 5 4 4 5 3 5 4 4 3 G| 1 3 2 1 2 1 3 2 1 4 4 3 4 4 T| 3 5 5 6 2 4 4 1 3 2 2 3 2 3
Pax-2	randomPax-2	A| 9 6 7 8 5 10 10 10 6 10 9 6 11 6 7 9 7 9 3 C| 8 9 9 7 8 6 9 7 8 6 7 7 10 6 11 12 5 10 8 G| 7 10 8 9 10 11 8 7 9 8 8 8 6 9 9 3 11 5 9 T| 8 7 8 8 9 5 5 8 9 8 8 11 5 11 5 8 9 8 12
Pax-3	randomPax-3	A| 8 6 13 9 7 8 5 12 8 5 4 11 7 C| 7 10 4 4 9 5 8 3 4 7 8 3 5 G| 6 7 7 7 5 8 5 5 6 3 8 7 7 T| 5 3 2 6 5 5 8 6 8 11 6 5 7
Pax-4a	randomPax-4a	A| 9 8 9 8 6 9 8 11 6 10 5 10 8 8 11 8 12 10 7 8 8 C| 14 11 16 13 17 12 13 9 15 9 13 13 17 17 8 14 11 11 14 14 14 G| 16 13 13 10 12 11 16 15 17 10 14 16 13 12 16 10 14 10 13 9 13 T| 4 11 5 12 8 11 6 8 5 14 11 4 5 6 8 11 6 12 9 12 8
Pax-5_M00143	randomPax-5_M00143	A| 3 2 1 1 0 2 1 1 1 1 2 1 0 2 1 0 2 2 2 1 1 1 0 1 0 1 2 0 C| 2 3 1 3 4 3 0 1 2 4 2 1 4 0 2 4 1 2 1 3 2 4 2 1 5 2 2 4 G| 2 1 3 2 2 2 5 4 2 1 3 4 1 2 1 2 2 2 2 2 2 0 4 4 2 2 3 3 T| 0 1 2 1 1 0 1 1 2 1 0 1 2 3 3 1 2 1 2 1 2 2 1 1 0 2 0 0
Pax-5_M00144	randomPax-5_M00144	A| 1 4 1 2 0 0 1 2 1 3 2 0 2 1 2 1 0 1 2 1 1 2 2 1 2 0 3 1 C| 1 0 1 1 1 2 1 1 1 0 1 1 2 0 2 2 2 2 1 2 2 2 0 1 0 3 1 3 G| 3 0 1 1 3 1 1 2 1 1 1 3 1 2 1 2 1 1 1 1 2 0 1 3 1 0 0 1 T| 0 1 2 1 1 2 2 0 2 1 1 1 0 2 0 0 2 1 1 1 0 1 2 0 2 2 1 0
Pax-6	randomPax-6	A| 16 11 21 15 13 13 10 14 10 16 14 19 16 15 10 13 14 15 9 14 11

		C| 10 12 9 13 12 7 15 13 8 8 8 10 9 8 10 7 5 8 8 6 13 G| 6 13 10 8 6 13 12 6 8 10 14 9 10 15 16 13 10 9 14 9 12 T| 15 11 7 11 16 14 10 14 21 13 11 9 12 9 11 14 18 15 16 18 11
POU2F1_M00136	randomPOU2F1_M00136	A| 16 16 19 16 13 16 16 7 17 17 17 17 10 12 13 C| 11 7 5 7 7 8 10 11 10 5 5 8 11 8 6 G| 6 7 8 7 7 3 6 12 4 5 7 7 7 9 6 T| 11 14 12 14 17 17 12 14 13 17 15 12 16 15 19
POU2F1_M00162	randomPOU2F1_M00162	A| 1 2 2 2 2 4 2 0 3 4 2 2 0 4 C| 0 2 2 1 0 1 2 1 0 1 2 2 2 1 G| 0 1 1 1 1 0 1 2 0 0 1 0 1 0 T| 5 1 1 2 3 1 1 3 3 1 1 2 3 1
PPAR-alpha	randomPPAR-alpha	A| 1 2 2 2 1 2 2 1 0 3 4 1 2 4 1 2 2 3 3 2 C| 1 3 1 0 4 2 0 1 5 2 1 1 1 2 1 1 3 2 3 0 G| 0 1 2 2 1 1 4 2 1 1 1 2 2 0 2 3 2 1 0 3 T| 5 1 2 3 1 2 1 3 1 1 1 3 2 1 3 1 0 1 1 2

Tabelle 35: Vergleich der bei den pflanzlichen Matrizen vorhandenen Kernsequenz bei zwei unterschiedlichen Suchmodi, einmal wurde bei Patser die Anzahl der ausgegebenen TFBSn pro Matrize approximativ auf 7500 eingestellt, das andere Mal wurde der von Patser automatisch bestimmte Mindestscore verwendet.

Transkriptions-faktor	**Patser-score bei ca. 7500 TFBSn**	**Anzahl TFBSn**	**Anzahl TFBSn mit Kernsequenz**	**Anzahl TFBSn beim Patser auto-score**	**Anzahl TFBSn mit Kernsequenz beim Patser auto-score**
ABF1	8,45	6998	5994	2095	2095
ABI4(1)	9,53	7352	7180	12830	10797
ABI4(2)	8,79	7552	6426	11955	9781
AG_M00151	8,24	7537	4944	41343	23711
AGL15	7,8	7537	7537	164867	161816

AGL3	8,06	7561	6514	59248	39527
AGP1	8,39	7700	7700	85589	78459
ALFIN1	8,68	7561	7561	546159	439199
ANT	6,9	7570	4081	294	234
ATHB1	8,11	8592	8592	66460	66460
ATHB5	7,8	7576	7266	6480	6334
ATHB9_M00417	6,35	7599	4496	291	268
AtMYB15	6,05	7503	5665	209	209
AtMYB77	8,39	7529	7529	17836	17822
AtMYB84	6,47	7591	5262	231	231
bZIP910(1)	7,4	7523	3422	345	345
bZIP910(2)	7,1	7594	3707	470	406
bZIP911(1)	6,63	7605	3568	123	123
bZIP911(2)	7,5	8033	3123	145	134
CBF	6,46	7466	7466	1337117	1337117
CDC5_M00361	7,9	7766	6891	11574	10119
DOF2_M00353	7,26	7682	7682	1840355	1703538
GAMYB	8	7547	7547	315722	300027
GT1	7,72	7553	7553	1439744	1040071
HAHB4	8,4	8644	4370	90825	45426
HVH21	8,17	7525	7525	526877	405865
ID1	8,95	7508	7426	156641	54182
MYB.PH3_M00218	7,82	7475	7347	8578	8340
MYB.PH3_M00219	8,02	7495	7112	7652	7211
O2	8,35	7551	7551	100666	52511
P	8,63	8275	4183	210035	94357
PCF2	8,07	7590	7354	35623	29907
PCF5	8,03	7583	7583	13554	13554
PIF3_M00434	7,56	7573	4034	918	825
PIF3(2)	7,17	7589	3977	834	798
RAV1_M00343	8,66	7590	7590	310764	310632
RAV1_M00344	8,64	7573	7573	229983	229054
TBP	8,15	7655	7647	199369	173356
TEIL	8,1	8450	8450	602300	593450
TGA1	8,73	7568	7568	52803	52238
TGA1a	8,05	7592	7592	81348	45553

ZAP1	7,86	7605	3746	4302	2116

Tabelle 36: Konsensus-Sequenz der Matrizen der pflanzlichen Transkriptionsfaktoren. Die Kernsequenz ist fett gedruckt und repräsentiert die konserviertesten Positionen der Matrize.

Transkriptionsfaktor	**Konsensus-Sequenz**
ABF1	AC**ACGT**GGC
ABI4(1)	CA**CCGC**CCCN
ABI4(2)	**CGGT**GCNNCC
AG_M00151	NTTWCCW**AA**WNNGGNAAN
AGL15	TTW**C**YWWWWWTRGNAA
AGL3	NNTN**CCA**WWWATAGNWNN
AGP1	NC**AGATC**TAN
ALFIN1	RN**GTG**GGGN
ANT	G**CAC**ANWTCCCNAKGT
ATHB1	CAATT**ATT**G
ATHB5	BSNNCAATT**ATT**G
ATHB9_M00417	NNNNGTA**ATGAT**TACNYBC
AtMYB15	CVAC**GGT**A**GGT**GG
AtMYB77	NRDYGRCR**GTT**RS
AtMYB84	NGGD**GGT**A**GGT**GS
bZIP910(1)	GATG**ACGT**GGCM
bZIP910(2)	GGRTGCTG**ACGT**
bZIP911(1)	GRTG**ACGT**GGCC
bZIP911(2)	GRTG**ACGT**GTAC
CBF	NNNNN**CAAT**NNNN
CDC5_M00361	NG**CT**CAGCGCN
DOF2_M00353	NNNW**AAAG**CNN
GAMYB	Y**AAC**SGMC
GT1	RNANN**GT**AAAAW
HAHB4	T**AAT**GATTTG
HVH21	NNYST**GAC**RSNK
ID1	T**TTGTC**GTTTT
MYB.PH3_M00218	WWAAAACS**GTTA**Y
MYB.PH3_M00219	NRRAGTTA**GTT**AS
O2	NHC**ACGT**GDN
P	ACCW**ACC**NNNN

PCF2	NNGGN**CCC**AC
PCF5	GTGGY**CC**CNN
PIF3_M00434	GKRGGM**CACGTG**RMSWCK
PIF3(2)	NNNVC**CACGT**GGNMVNN
RAV1_M00343	NNGCAA**CA**KAWN
RAV1_M00344	NNCAC**CT**GRNNN
TBP	NC**TATA**AATANN
TEIL	AT**G**WAYCT
TGA1	NNNV**TGAC**GTTA
TGA1a	SNC**ACGT**GNS
ZAP1	**TTGAC**CGAGY

Tabelle 37: Durchschnittlicher AT-Gehalt im intergenischen Bereich, in der 5'UTR, in Introns, Exons sowie der 3'UTR. Des Weiteren ist dieser für die erste 5'UTR, die ersten fünf Intons, die ersten vier Exons sowie die erste 3'UTR angegeben. Der durchschnittliche AT-Gehalt im gesamten Genom beträgt 64,13 %.

Bereich	**AT-Gehalt [%]**
intergenisch	68,37
5UTR	61,84
INTRON	67,59
EXON	56,39
3UTR	66,47
5UTR1	61,89
INTRON1	68,13
INTRON2	68,18
INTRON3	67,75
INTRON4	67,38
INTRON5	67,18
EXON1	56,47
EXON2	56,1
EXON3	56,39
EXON4	56,31
3UTR1	66,71

Tabelle 38: Eingestellter Mindest-score bei den Zufallsmatrizen der pflanzlichen Matrizen incl. der Anzahl der detektierten TFBSn

Name der Matrize	Mindest-score bei ca. 7500 TFBSn	Anzahl TFBSn
randomABF1	5,9	7462
randomABI4_1	8,15	7541
randomABI4_2	7,43	7599
randomAG	2,69	7364
randomAGL15	3,69	7443
randomAGL3	2,5	7479
randomAGP1	3,67	8083
randomALFIN1	6,46	7786
randomANT	5,98	7522
randomATHB1	3,96	10173
randomATHB5	3,36	7552
randomATHB9	4,67	7548
randomAtMYB15	6,16	7466
randomAtMYB77	4,23	7541
randomAtMYB84	7,36	7535
randombZIP910_1	6,42	7401
randombZIP910_2	6,14	7516
randombZIP911_1	7,4	7458
randombZIP911_2	5,59	7563
randomCBF	1,52	8288
randomCDC5	6,55	7512
randomDOF2	3,13	7579
randomGAMYB	4,59	8961
randomGT1	3,1	7576
randomHAHB4	4,43	8452
randomHVH21	5,07	7457
randomID1	4,01	7984
randomMYB.PH3_1	2,96	7695
randomMYB.PH3_2	3,22	7835
randomO2	5,75	7331
randomP	4,17	7496
randomPCF2	7,42	7693
randomPCF5	7,07	7839

randomPIF3_1	7,76	7430
randomPIF3_2	7,64	7513
randomRAV1_1	3,47	7607
randomRAV1_2	5,14	7456
randomTBP	3,07	8313
randomTEIL	1,83	9346
randomTGA1	4,68	7313
randomTGA1a	6,43	7508
randomZAP1	4,3	8108

Tabelle 39: Eingestellter Mindest-score bei den nichtpflanzlichen Matrizen incl. der Anzahl der TFBSn nach einer Redundanzanalyse und dem Entfernen der doppelten genomischen Positionen

Transkriptionsfaktor	Mindest-score bei ca. 7500 TFBSn	Anzahl TFBSn
abaA	9,06	7554
AP-2alpha	8,31	7522
AREB6_M00414	8,11	7552
AREB6_M00415	7,86	9999
Brachyury	1,69	7555
c-Ets-1	8,51	7519
c-Ets-154	8,03	7565
COUP	8,47	7539
CP2	8,53	7518
Dl	9,03	7650
E2F	8,08	7567
E74A	8,65	7528
FOXA4a	8,07	7482
GAL4	7,77	7577
GCM	7,7	7569
HAP3	8,95	7538
LAC9	7,51	7584
NF-kappaB1	8,3	7536
Pax-2	7,37	7533
Pax-3	7,57	7571
Pax-4a	8,8	7442
Pax-5_M00143	8,55	7529
Pax-5_M00144	8,02	7558

Pax-6	7,95	7550
POU2F1_M00136	8,13	7515
POU2F1_M00162	7,75	7451
PPAR-alpha	7,36	7581

Tabelle 40: Eingestellter Mindest-score bei den Zufallsmatrizen der nichtpflanzlichen Matrizen incl. der Anzahl der detektierten TFBSn

Name der Matrize	**Mindest-score bei ca. 7500 TFBSn**	**Anzahl TFBSn**
randomabaA_M00027	6,72	7552
randomAP-2alpha_M00189	7,43	7618
randomAREB6_M00414	6,2	7617
randomAREB6_M00415	3,8	7525
randomBrachyury_M00150	5,19	7556
randomc-Ets-1_M00032	4,39	7049
randomc-Ets-154_M00074	3,84	7436
randomCOUP_M00158	5,15	7442
randomCP2_M00072	6,15	7031
randomDl_M00120	5	8119
randomE2F_M00024	6,72	7484
randomE74A_M00016	5,53	7459
randomFOXA4a_M00267	5,6	7045
randomGAL4_M00049	8,09	7551
randomGCM_M00270	5,48	7555
randomHAP3_M00288	6,18	7466
randomLAC9_M00207	6,67	7483
randomNF-kappaB1_M00194	6	7177
randomPax-2_M00098	5,56	7635
randomPax-3_M00360	4,43	7412
randomPax-4a_M00373	7,81	7546
randomPax-5_M00143	9,82	7536
randomPax-5_M00144	8,32	7562
randomPax-6_M00097	4,41	7494
randomPOU2F1_M00136	2,71	7895
randomPOU2F1_M00162	5,71	7137
randomPPAR-alpha_M00242	7,15	7559

Tabelle 41: Durchschnittliche Anzahl der Transkriptionsfaktor-Bindungsstellen sowie der Zufallspositionen im intergenischen Bereich, sowie in der 5'UTR, in Introns, Exons und der 3'UTR, angegeben in Prozenten.

Versuchsgruppe	Anzahl TFBSen [%] intergenisch	Anzahl TFBSen [%] 5'UTR	Anzahl TFBSen [%] Intron	Anzahl TFBSen [%] Exon	Anzahl TFBSen [%] 3'UTR
Zufallspositionen	43,85	1,77	15,85	35,24	3,29
pflanzliche Matrizen	40,02	1,59	11,02	44,92	2,45
Zufallsmatrizen pflanzlich	37,68	2,25	7,13	51,04	1,9
nichtpflanzliche Matrizen	32,13	1,85	10,52	53,2	2,31
Zufallsmatrizen nichtpflanzlich	28,8	2,2	5,83	61,58	1,58

Tabelle 42: Durchschnittliche Anzahl der Transkriptionsfaktor-Bindungsstellen sowie der Zufallspositionen in den ersten fünf Introns und den ersten vier Exons angegeben in Prozenten.

Bereich innerhalb von Genen	Anzahl Zufallspositionen [%]	Anzahl TFBSn pflanzlicher Matrizen [%]	Anzahl TFBSn Zufallsmatrizen pflanzlich [%]	Anzahl TFBSn nichtpflanzlicher Matrizen [%]	Anzahl TFBSn Zufallsmatrizen nichtpflanzlich [%]
Intron1	4,35	3,08	2,02	2,74	1,67
Intron2	2,57	1,75	1,21	1,59	0,94
Intron3	1,79	1,22	0,77	1,13	0,63
Intron4	1,38	0,95	0,66	0,9	0,54
Intron5	1,06	0,73	0,43	0,75	0,36
Exon1	17,43	23,64	33,42	27,71	39,03
Exon2	4,4	5,7	5,68	6,63	7,06
Exon3	3,07	3,69	3,22	4,37	4,06
Exon4	2,22	2,55	1,97	3,04	2,53

Tabelle 43: Abweichungsfaktoren der pflanzlichen Matrizen im Vergleich zu den Zufallspositionen im intergenischen Bereich sowie in der 5'UTR und den ersten vier Exons.

Matrize	intergenisch	5UTR1	Exon1	Exon2	Exon3	Exon4
ABF1	1,26	0,69	1,24	1,22	0,82	0,77
ABI4(1)	0,31	2,11	3,41	2,06	1,37	0,85
ABI4(2)	0,44	1,08	2,52	2,22	1,83	1,58
AG_M00151	1,07	1,18	0,86	0,83	0,83	0,86
AGL15	1,36	0,97	0,46	0,46	0,46	0,48
AGL3	1,18	1,08	0,68	0,65	0,61	0,74
AGP1	0,79	2,29	1,36	1,18	0,99	0,98
ALFIN1	0,64	0,66	2,16	1,7	1,55	1,44
ANT	0,73	1,41	1,45	1,47	1,51	1,46
ATHB1	1,44	0,58	0,42	0,39	0,52	0,49
ATHB5	1,07	0,78	0,88	0,96	1,04	1,15
ATHB9_M00417	0,75	0,66	1,38	1,37	1,51	1,4
AtMYB15	0,64	0,97	1,75	1,67	1,68	1,65
AtMYB77	0,69	0,96	1,66	1,69	1,54	1,69
AtMYB84	0,63	0,95	1,89	1,65	1,5	1,42
bZIP910(1)	0,89	0,71	1,61	1,52	1,39	1,27
bZIP910(2)	0,73	0,78	1,66	1,8	1,7	1,33
bZIP911(1)	0,84	0,68	1,71	1,56	1,31	1,37
bZIP911(2)	0,88	0,73	1,42	1,38	1,29	1,21
CBF	0,89	0,45	1,32	1,53	1,18	1,4
CDC5_M00361	0,43	0,93	2,06	2,03	2,03	1,95
DOF2_M00353	0,91	1,07	1,1	1,13	1,16	1,1
GAMYB	0,71	0,96	1,95	1,65	1,46	1,56
GT1	1,27	0,76	0,53	0,47	0,57	0,52
HAHB4	1,35	0,53	0,53	0,51	0,48	0,45
HVH21	0,68	1,12	1,94	1,73	1,39	1,54
ID1	1,2	1,34	0,68	0,59	0,7	0,65
MYB.PH3_M00218	1,08	0,73	0,92	1,04	0,87	0,93
MYB.PH3_M00219	1,08	0,76	0,76	0,69	0,67	0,62
O2	1,26	0,79	1,23	1,21	0,88	0,74
P	0,78	0,95	1,32	1,36	1,45	1,24
PCF2	1,12	0,49	1,21	1,21	1,19	1,16
PCF5	0,86	0,48	1,4	1,54	1,4	1,77

PIF3_M00434	1,01	0,63	1,46	1,44	1,37	1,22
PIF3(2)	1	0,58	1,46	1,53	1,31	1,1
RAV1_M00343	0,61	0,85	1,42	1,58	1,68	1,52
RAV1_M00344	0,39	0,77	1,84	1,92	1,88	1,74
TATABP	1,6	1,01	0,26	0,29	0,26	0,32
TEIL_M00502	0,74	0,29	1,03	1,35	1,38	1,44
TGA1	0,89	1,1	1,71	1,36	1,32	0,98
TGA1a	1,16	0,99	1,33	1,25	0,97	0,85
ZAP1	0,9	0,79	1,29	1,38	1,51	1,29

Tabelle 44: Abweichungsfaktoren der pflanzlichen Matrizen im Vergleich zu den Zufallspositionen in den ersten fünf Introns und der ersten 3'UTR.

Matrize	**Intron1**	**Intron2**	**Intron3**	**Intron4**	**Intron5**	**3UTR1**
ABF1	0,35	0,39	0,26	0,17	0,21	0,58
ABI4(1)	0,16	0,11	0,07	0,08	0,04	0,28
ABI4(2)	0,3	0,17	0,29	0,16	0,17	0,37
AG_M00151	1,3	1,12	1,14	0,9	0,75	0,99
AGL15	1,49	1,25	1,17	1,27	1,05	1,08
AGL3	1,4	1,08	1,02	1,27	0,99	0,91
AGP1	1,86	0,83	0,66	0,51	0,54	0,63
ALFIN1	0,26	0,32	0,38	0,36	0,44	0,52
ANT	0,8	0,72	0,64	0,53	0,74	0,68
ATHB1	1,18	1,07	0,96	1,07	0,89	1,13
ATHB5	1,02	0,82	0,75	0,85	0,81	0,99
ATHB9_M00417	0,89	0,98	0,73	0,87	0,74	0,82
AtMYB15	0,49	0,62	0,61	0,71	0,58	0,56
AtMYB77	0,4	0,5	0,57	0,63	0,64	0,57
AtMYB84	0,42	0,51	0,46	0,65	0,65	0,57
bZIP910(1)	0,31	0,36	0,39	0,28	0,26	0,59
bZIP910(2)	0,41	0,45	0,53	0,64	0,47	0,48
bZIP911(1)	0,32	0,39	0,3	0,33	0,36	0,6
bZIP911(2)	0,56	0,61	0,61	0,69	0,67	0,86
CBF	0,57	0,66	0,52	0,44	0,49	0,8
CDC5_M00361	0,35	0,31	0,31	0,38	0,4	0,52
DOF2_M00353	1,01	0,92	0,97	0,82	1,02	1,03

GAMYB	0,31	0,35	0,33	0,33	0,33	0,55
GT1	1,12	1,12	1,37	1,17	1,46	1,21
HAHB4	1,17	1,14	1,32	1,09	1,21	1,04
HVH21	0,31	0,33	0,39	0,43	0,46	0,55
ID1	1,3	1,08	1,1	0,99	1,09	1,3
MYB.PH3_M00218	0,84	0,96	1,09	1,04	1,06	0,98
MYB.PH3_M00219	1,24	1,43	1,52	1,57	1,9	0,99
O2	0,43	0,42	0,3	0,22	0,21	0,6
P	0,67	0,86	0,91	0,91	1,01	0,93
PCF2	0,37	0,36	0,35	0,36	0,24	0,49
PCF5	0,39	0,5	0,4	0,41	0,36	0,48
PIF3_M00434	0,31	0,31	0,32	0,29	0,37	0,55
PIF3(2)	0,32	0,3	0,38	0,32	0,31	0,55
RAV1_M00343	0,77	0,83	0,92	1,02	1,28	0,84
RAV1_M00344	0,52	0,53	0,81	0,84	0,76	0,57
TATABP	1	1,15	1,01	0,86	0,89	0,97
TEIL_M00502	1,03	1,29	1,32	1,59	1,47	0,99
TGA1	0,48	0,39	0,41	0,51	0,45	0,61
TGA1a	0,46	0,34	0,32	0,34	0,2	0,62
ZAP1	0,64	0,57	0,56	0,66	0,65	0,56

Tabelle 45: 5UTR1, Auswahl der niedrigsten 10 P-Werte und Gegenüberstellung der Expression des Transkriptionsfaktors und der entsprechenden putativen Zielgene in der ersten 5'UTR. Für die Expressionsdaten der putativen Zielgene wurde Genevestigator eingesetzt. Die Daten bezüglich der Expression des Transkriptionsfaktors stammen entweder aus Genevestigator, wenn der Faktor ebenfalls aus *Arabidopsis thaliana* stammt, oder aus TRANSFAC©. Übereinstimmungen der Expressionsdaten der putativen Zielgene und des Transkriptionsfaktors sind fett gedruckt.

TF	**Distanz**	**Stimulus**	**Development**	**Anatomy**	**TF Expression**
AGP1	52	keine Gemeinsam-keiten	Expression: ausgereifte Schoten; Keimling, bolting	Expression Stängel; unterschiedlich Expression: Zellsuspension; schwache Expression:	Expression: Blätter, suspensions-kultivierte Zellen, **Blüte**, **Wurzel**

				Blütenstand, Keimling, Austrieb des Pflanzen-sprosses, verschiedene **Wurzel**-Bereiche	
ABI4(1)	96	keine Gemeinsam-keiten	Expression: ausgereifte Schoten; ausgekeimte **Samen**	Expression: Stängel; Stengelblatt	pos: sich entwickelnde **Samen,** leicht erhöht bei Samenreifung; vegetatives Gewebe, besonders im vaskulären Gewebe; neg: nicht induzierbar durch Trockenheit, Salz, Glucose oder Kältestress
ABI4(1)	44	keine Gemeinsam-keiten	Expression: ausgereifte Schoten	Expression: Stängel; unterschiedliche Expression: Wurzelspitzen	pos: sich entwickelnde Samen, leicht erhöht bei Samenreifung; vegetatives Gewebe, besonders im vaskulären Gewebe; neg: nicht induzierbar durch Trockenheit, Salz, Glucose oder Kältestress

ABI4(1)	32	upreguliert: Chemikalie Norflurazon; 4 von 5 rot: Isoxaben	Expression: ausgereifte Schoten; ausgekeimte **Samen**, bolting	Expression: Keimwurzel, Austrieb des Pflanzen-sprosses, Wurzelkompo-nenten; unterschiedlich Expression, aber teilweise intensiv: Callus, Zellsuspension, Samen, Stängel, Wurzelspitze	pos: sich entwickelnde **Samen**, leicht erhöht bei Samenreifung; vegetatives Gewebe, besonders im vaskulären Gewebe; neg: nicht induzierbar durch Trockenheit, Salz, Glucose oder Kältestress
ABI4(1)	26	unterschiedlich reguliert: Seneszenz	Expression: ausgereifte Schoten	Expression: Stängel	pos: sich entwickelnde Samen, leicht erhöht bei Samenreifung; vegetatives Gewebe, besonders im vaskulären Gewebe; neg: nicht induzierbar durch Trockenheit, Salz, Glucose oder Kältestress
ANT	79	unterschiedlich reguliert: Chemikalie Isoxaben	Expression: ausgereifte Schoten	Expression: Stängel	keine Gemeinsam-keiten
AtMYB77	70	upreguliert: *A.tumefaciens*; downreguliert:	Expression: **ausgereifte Schoten**,	Expression: Stängel, Wurzeln (und	Expression: **ausgereifte Schoten,**

		Chemikalie Cycloheximid, Seneszenz, osmotischem Stress; unterschiedlich reguliert: Chemikalie Syringolin	Keimling, keine Expression: Pollen	Komponenten), **Hypocotyl**, Keimwurzel, Blütenstand (Schote, **Samen**), keine Expression: **Pollen**	**Hypocotyl, Samen;** keine Expression: **Pollen**
ID1	7	downreguliert: 2,4-Dichlorophenoxylat, erhöhte CO_2 Konzentration; unterschiedlich: *B. cinerea*, Nematoden, *P. infestans*, *P.syringae*, Licht intensives Licht 2	keine Expression: ausgekeimte Samen, bolting; unterschiedliche Expression: ausgereifte Schoten	keine Expression: Blütenstands-konoten, Wurzelspitze, seitliche Wurzelkappe; unterschiedliche Expression: Stängel	Expression: anwesend während des ganzen postembryonalen Lebens der sporophyten Pflanze in: unreifem Blatt, letztem Blatt, eingekapselte unreife männliche Blüte; keine Expression: reifes Blatt, Spitze transformiert in männliches Blüten-Primordium, Wurzel, unreife männlicher Blüte
ID1	2	unterschiedlich reguliert: Chemikalie Cycloheximid, Hormon BL/H3BO3, Stress: Wärme	Expression: ausgereifte Schoten	Expression: Stängel	Expression: anwesend während des ganzen postembryonalen Lebens der sporophyten

					Pflanze in: unreifem Blatt, letztes Blatt, eingekapselte unreife männliche Blüte; keine Expression: reifes Blatt, Spitze transformiert in männliches Blüten-Primordium, Wurzel, unreife männliche Blüte
TATABP	92	unterschiedlich reguliert: *A.tumefaciens*, *B. cinerea, P. syringae,* Chemikalie Isoxaben, Ozon, ABA, Hormon BL/H_3BO_3, Kältestress; rot: Lichtintensität	unterschiedlich Expression: ausgereifte Schoten	keine Expression: Hypocotyl und Komponenten; unterschiedliche Expression: Stängel	ubiquitär

Tabelle 46: Intron1: Auswahl der niedrigsten 10 P-Werte und Auflistung der Expression der entsprechenden putativen Zielgene. Für die Expressionsdaten der putativen Zielgene wurde Genevestigator eingesetzt. Im Fall von O2 und TGA1a langen die putativen Bindungsstellen in je 14 rezeptorähnlichen Protein Kinasen, welche identisch sind. Beide Matrizen erkennen offensichtlich in diesen Fällen dieselben genomischen Positionen, worauf auch die Distanz hindeutet.

TF	**Distanz**	**Stimulus**	**Development**	**Anatomy**
AG_ M00151	4	keine Gemeinsamkeiten	keine Gemeinsamkeiten, aber Expression (25 von 26): ausgereifte Schoten	keine Gemeinsamkeiten, aber Expresssion (23 von 26): Stängel

AGL15	3	keine Gemeinsamkeiten	keine Gemeinsamkeiten	keine Gemeinsamkeiten
ATHB9_ M00417	14	keine Gemeinsamkeiten	keine Gemeinsamkeiten	keine Gemeinsamkeiten
GT1	6	keine Gemeinsamkeiten	keine Gemeinsamkeiten	keine Gemeinsamkeiten
MYBPH3_ M00219	0	unterschiedlich reguliert: Lichtintensität	keine Gemeinsamkeiten	keine Gemeinsamkeiten
O2	65	1 Arrayelement für alle: durch Chemikalien, Stress, Licht, Nährstoffe, Hormone, biotischen Stress rot bzw. grün	1 Arrayelement für alle: überall keine Expression	1 Arrayelement für alle: Expression: Staubgefäße, Pollen
P	5	keine Gemeinsamkeiten	keine Gemeinsamkeiten	keine Gemeinsamkeiten
RAV1_ M00344	4	keine Gemeinsamkeiten	keine Gemeinsamkeiten	keine Gemeinsamkeiten
TATABP	10	unterschiedlich reguliert: Chemikalie Cycloheximid	Expression: ausgereifte Schoten	Expression: Stängel
TGA1a	65	1 Arrayelement für alle: durch Chemikalien, Stress, Licht, Nährstoffe, Hormone, biotischen Stress upreguliert bzw. downreguliert	1 Arrayelement für alle: überall keine Expression	1 Arrayelement für alle: Expression: Staubgefäße, Pollen

Tabelle 47: Intron2: Auswahl der niedrigsten 10 P-Werte und Auflistung der Expression der entsprechenden putativen Zielgene. Für die Expressionsdaten der putativen Zielgene wurde Genevestigator eingesetzt. Expressionsdaten bezüglich des Transkriptionsfaktors wurden mittels TRANSFAC© festgestellt.

TF	**Distanz**	**Stimulus**	**Development**	**Anatomy**	**TF Expression**
AG_	4	keine	keine	keine	kein

M00151		Gemeinsamkeiten	Gemeinsamkeiten	Gemeinsamkeiten	Vergleich notwendig
AGL15	3	unterschiedlich reguliert: Seneszenz	keine Gemeinsamkeiten	keine Gemeinsamkeiten	kein Vergleich notwendig
ATHB9_ M00417	14	unterschiedlich reguliert: Chemikalie Syringolin	keine Gemeinsamkeiten	keine Gemeinsamkeiten	kein Vergleich notwendig
AtMYB15	7	unterschiedlich reguliert: Wärme2	keine Gemeinsamkeiten	keine Gemeinsamkeiten	kein Vergleich notwendig
GT1	6	keine Gemeinsamkeiten	keine Gemeinsamkeiten	keine Gemeinsamkeiten	kein Vergleich notwendig
MYBPH3_ M00219	0	unterschiedlich reguliert: *A. tumefaciens, B. cenerea,* Nematoden, Chemikalie Cycloheximid, Isoxaben, Norflurazon, Ozon, Syringolin, Seneszenz (stark!), Anoxie, Wärme	Expression: ausgereifte Schoten	Expression: Stängel	Expression: Blüte; keine Expression: Blätter und Stängel, Kelchblätter, Wurzeln
P	5	unterschiedlich reguliert: Chemikalie Cycloheximid	Expression: ausgereifte Schoten, junge Rosette	Expression: Stängel	Flavonoid-Metabolismus
RAV1_ M00344	4	keine Gemeinsamkeiten	keine Gemeinsamkeiten	keine Gemeinsamkeiten	kein Vergleich notwendig
TATABP	10	unterschiedlich reguiert:	Expression: junge Blüte, sich	Expression Rosette,	ubiquitär

		Chemikalie Cycloheximid, Isoxaben, Syringolin, Hormon BL/H3BO3, wenig-N--Glucose, Seneszenz, Kälte, Wärme, Hypoxie, Osmose, Salz	entwicklende Blüte	ausgewachsenes Blatt, seitliche Wurzel	
TEIL_ M00502	5	keine Gemeinsamkeiten	keine Gemeinsamkeiten	keine Gemeinsamkeiten	kein Vergleich notwendig

Tabelle 48: Intron3: Auswahl der niedrigsten 10 P-Werte und Auflistung der Expression der entsprechenden putativen Zielgene. Für die Expressionsdaten der putativen Zielgene wurde Genevestigator eingesetzt. Expressionsdaten bezüglich des Transkriptionsfaktors wurden mittels TRANSFAC© sowie Genevestigator, wenn der Transkriptionsfaktor aus *Arabidopsis thaliana* stammt, festgestellt. Übereinstimmungen der Expressionsdaten der putativen Zielgene und des Transkriptionsfaktors sind fett gedruckt.

TF	**Distanz**	**Stimulus**	**Development**	**Anatomy**	**TF-Expression**
AG_ M00151	4	unterschiedlich reguliert: Wärme	keine Gemeinsamkeiten	keine Gemeinsamkeiten	kein Vergleich notwendig
AGL15	3	unterschiedlich reguliert: Wärme, Hypoxie	keine Gemeinsamkeiten	Expression: Elongationszone, Leitbündel	keine Gemeinsam-keiten
AtMYB15	7	keine Gemeinsamkeiten	keine Gemeinsamkeiten	keine Gemeinsamkeiten	kein Vergleich notwendig
AtMYB84	7	downreguliert (13 von 14): Chemikalie 6-	Expression: **ausgereifte Schoten**	keine Gemeinsamkeiten	Expression: **ausgereifte Schoten**

		Benzyl-Adenin; unterschiedlich reguliert: Cycloheximid			
bZIP911(2)	0	unterschiedlich reguliert: Wärme2	Expression: ausgereifte Schoten	Expression: Stängel	Expression: Blüten Wurzeln
GT1	6	keine Gemeinsamkeiten	keine Gemeinsamkeiten	keine Gemeinsamkeiten	kein Vergleich notwendig
MYBPH3_ M00219	0	downreguliert (5 von 6): Seneszenz	Expression: ausgereifte Schoten	Expression: Stängel	Expression: Blüte; keine Expression: Blätter und Stängel, Kelchblätter, Wurzeln
P	5	keine Gemeinsamkeiten	keine Gemeinsamkeiten	keine Gemeinsamkeiten	kein Vergleich notwendig
RAV1_ M00344	4	keine Gemeinsamkeiten	keine Gemeinsamkeiten	keine Gemeinsamkeiten	kein Vergleich notwendig
TEIL_ M00502	5	keine Gemeinsamkeiten	keine Gemeinsamkeiten	keine Gemeinsamkeiten	kein Vergleich notwendig

Tabelle 49: Intron4: Auswahl der niedrigsten 10 P-Werte und Auflistung der Expression der entsprechenden putativen Zielgene. Für die Expressionsdaten der putativen Zielgene wurde Genevestigator eingesetzt. Expressionsdaten bezüglich des Transkriptionsfaktors wurden mittels TRANSFAC© sowie Genevestigator, wenn der Transkriptionsfaktor aus *Arabidopsis thaliana* stammt, festgestellt.

TF	**Distanz**	**Stimulus**	**Development**	**Anatomy**	**TF-Expresssion**
AG_	4	keine	keine	keine	kein

M00151		Gemeinsamkeiten	Gemeinsamkeiten	Gemeinsamkeiten	Vergleich notwendig
AGL15	3	unterschiedlich reguliert: Licht_2	keine Gemeinsamkeiten	keine Gemeinsamkeiten	kein Vergleich notwendig
AGL3	4	unterschiedlich reguliert: (8 von 9): Chemikalie Syringolin	keine Gemeinsamkeiten	keine Gemeinsamkeiten	kein Vergleich notwendig
AtMYB15	7	keine Gemeinsamkeiten	13 von 14: exprimiert: ausgereifte Schoten	keine Gemeinsamkeiten	kein Vergleich notwendig
AtMYB84	7	unterschiedlich reguliert: Chemikalie Syringolin, Wärme_2	Expression: ausgereifte Schoten	Expression: Stängel	keine Gemeinsam-keiten
GT1	6	unterschiedlich reguliert: Seneszenz, Wärme	keine Gemeinsamkeiten	keine Gemeinsamkeiten	kein Vergleich notwendig
MYBPH3_ M00219	52	downreguliert: AgNO3, AVG, Isoxaben, Seneszenz, unterschiedlich reguliert: Cycloheximid, Hydogenperoxid, NPA, Ozon, Hormon MJ, Anoxie, Kälte, Wärme; rot: Hormon ACC, UV-Filter	keine Gemeinsamkeiten	keine Gemeinsamkeiten	Expression: Blüte; keine Expression: Blätter und Stängel, Kelchblätter, Wurzeln; Stimuli unbekannt
P	5	keine	keine	keine	kein

		Gemeinsamkeiten	Gemeinsamkeiten	Gemeinsamkeiten	Vergleich notwendig
RAV1_ M00344	4	keine Gemeinsamkeiten	keine Gemeinsamkeiten	keine Gemeinsamkeiten	kein Vergleich notwendig
TEIL_ M00502	5	keine Gemeinsamkeiten	keine Gemeinsamkeiten	keine Gemeinsamkeiten	kein Vergleich notwendig

Tabelle 50: Intron5: Auswahl der niedrigsten 10 P-Werte und Auflistung der Expression der entsprechenden putativen Zielgene. Für die Expressionsdaten der putativen Zielgene wurde Genevestigator eingesetzt. Expressionsdaten bezüglich des Transkriptionsfaktors wurden mittels TRANSFAC© sowie Genevestigator, wenn der Transkriptionsfaktor aus *Arabidopsis thaliana* stammt, festgestellt. Übereinstimmungen der Expressionsdaten der putativen Zielgene und des Transkriptionsfaktors sind fett gedruckt.

TF	**Distanz**	**Stimulus**	**Development**	**Anatomy**	**TF-Expresssion**
AG_ M00151	4	unterschiedlich reguliert: Seneszenz, Kälte	Expression: ausgereifte Schoten	Expression: Stängel	keine Gemeinsam-keiten
AGL15	3	upreguliert: Glucose; unterschiedlich reguliert: Seneszenz, Wärme	Expression: **ausgereifte Schoten**	Expression: Stängel, Wurzelspitze	Expression: **ausgereifte Schoten**
AtMYB15	7	unterschiedlich reguliert: Isoxaben, Glucose, Seneszenz, Wärme	Expression: ausgereifte Schoten	Expression: Stängel	keine Gemeinsam-keiten
AtMYB84	7	downreguliert: K(-); unterschiedlich reguliert: Dichlorophenoxalat, Isoxaben, Norflurazon, Syringolin,	Expression: **ausgereifte Schoten, Hypocotyl, Blüte**, Blütenstandsknoten, Austrieb des Pflanzensprosses,	Expression: **Blüte**, Staubgefäße	Expression: **ausgereifte Schoten, Blüte, Hypocotyl**

		Seneszenz, Hypoxie;	rosette		
GT1	6	keine Gemeinsamkeiten	keine Gemeinsamkeiten	keine Gemeinsamkeiten	kein Vergleich notwendig
MYBPH3_ M00219	0	unterschiedlich reguliert: Dichlorophenoxalat, Wärme (8 von 9)	Expression: ausgereifte Schoten	Expression: Stängel	Expression: Blüte; keine Expression: Blätter und Stängel, Kelchblätter, Wurzeln;
MYBPH3_ M00219	12	keine Gemeinsamkeiten	Expression: ausgereifte Schoten	keine Gemeinsamkeiten	kein Vergleich notwendig
P	5	keine Gemeinsamkeiten	keine Gemeinsamkeiten	Expression: Stängel	Involviert in Flavonoid-Metabolismus
RAV1_ M00344	4	keine Gemeinsamkeiten	Expression: ausgereifte Schoten	Expression Stängel	keine Gemeinsam-keiten
TEIL_ M00502	5	keine Gemeinsamkeiten	keine Gemeinsamkeiten	keine Gemeinsamkeiten	kein Vergleich notwendig

Tabelle 51: Exon1: Auswahl der niedrigsten 10 P-Werte und Auflistung der Expression der entsprechenden putativen Zielgene. Für die Expressionsdaten der putativen Zielgene wurde Genevestigator eingesetzt. Expressionsdaten bezüglich des Transkriptionsfaktors wurden mittels TRANSFAC© sowie Genevestigator, wenn der Transkriptionsfaktor aus *Arabidopsis thaliana* stammt, festgestellt.

TF	**Distanz**	**Stimulus**	**Development**	**Anatomy**	**TF-Expression**
ABI4(1)	29	keine Gemeinsam-keiten	keine Gemeinsamkeiten	keine Gemeinsam-keiten	kein Vergleich notwendig

ABI4(1)	41	unterschiedlich reguliert: Wärme	Expression: ausgereifte Schoten	keine Gemeinsam-keiten	pos: sich entwickelnde Samen, leicht erhöht bei Samenreifung; vegetatives Gewebe, besonders im vaskulären Gewebe; neg: nicht induzierbar durch Trockenheit, Salz, Glucose oder Kältestress
ABI4(1)	53	keine Gemeinsam-keiten	keine Gemeinsamkeiten	keine Gemeinsam-keiten	kein Vergleich notwendig
ABI4(1)	74	23 von 24 unterschiedlich reguliert: Syringolin	Expression: ausgereifte Schoten	Expression: Stängel	pos: sich entwickelnde Samen, leicht erhöht bei Samenreifung; vegetatives Gewebe, besonders im vaskulären Gewebe; neg: nicht induzierbar durch Trockenheit, Salz, Glucose oder Kältestress

ABI4(1)	62	23 von 24 unterschiedlich reguliert: Syringolin	keine Gemeinsamkeiten	keine Gemeinsam-keiten	kein Vergleich notwendig
ABI4(1)	50	keine Gemeinsam-keiten	Expression: ausgereifte Schoten	keine Gemeinsam-keiten	pos: sich entwickelnde Samen, leicht erhöht bei Samenreifung; vegetatives Gewebe, besonders im vaskulären Gewebe; neg: nicht induzierbar durch Trockenheit, Salz, Glucose oder Kältestress
ABI4(1)	104	keine Gemeinsam-keiten	keine Gemeinsamkeiten	keine Gemeinsam-keiten	kein Vergleich notwendig
ATHB5	7	unterschiedlich reguliert: Wärme	keine Gemeinsamkeiten	keine Gemeinsam-keiten	kein Vergleich notwendig
bZIP910(2)	112	unterschiedlich reguliert: Saccharose	Expression: ausgereifte Schoten	keine Gemeinsam-keiten	pos: Blüten, Wurzeln, Blätter
RAV1_ M00344	464	downreguliert: Isoxaben; unterschiedlich reguliert Salicylsäure, Seneszenz, Wärme;	Expression: ausgereifte Schoten	keine Gemeinsam-keiten	keine Gemeinsam-keiten

		upreguliert: wenig Nitrat			

Tabelle 52: Exon2: Auswahl der niedrigsten 10 P-Werte und Auflistung der Expression der entsprechenden putativen Zielgene. Für die Expressionsdaten der putativen Zielgene wurde Genevestigator eingesetzt. Expressionsdaten bezüglich des Transkriptionsfaktors wurden mittels TRANSFAC© sowie Genevestigator, wenn der Transkriptionsfaktor aus *Arabidopsis thaliana* stammt, festgestellt. Übereinstimmungen der Expressionsdaten der putativen Zielgene und des Transkriptionsfaktors sind fett gedruckt.

TF	**Distanz**	**Stimulus**	**Development**	**Anatomy**	**TF-Expression**
AGP1	5	unterschiedlich reguliert: Seneszenz, Wärme, Osmose	Expression: ausgereifte Schoten	keine Gemeinsam-keiten	Expression: Blätter, suspensionskultivierte Zellen, Blüte, Wurzel
bZIP910(2)	162	1 Arrayelement für alle: durch Chemikalien, Stress, Licht, Nährstoffe, Hormone, biotischen Stress upreguliert bzw. downreguliert	1 Array-element überall keine Expression	1 Array-element Expression: Staubgefäße, Pollen	pos: Blüten, Wurzeln, Blätter
HVH21	1	downreguliert AgNO3; unterschiedlich reguliert: *P. infestans*, Cycloheximid, Ozon, Kälte	keine Gemeinsam-keiten	Expression: Wurzel Haarzone	pos: Lemma, Spelze, vaskuläres Bündel , basale Region inclusive des SAM
ID1	258	1 Arrayelement	1 Array-element	1 Array-element	pos: während des ganzen

		für alle: durch Chemikalien, Stress, Licht, Nährstoffe, Hormone, biotischen Stress upreguliert bzw. downreguliert	überall keine Expression	Expression: Staubgefäße, Pollen	postembryonalen Lebens der sporophyten Pflanze im: unreifen Blatt, letztem Blatt, eingekapselter unreifer männlicher Blüte; neg: reifen Blatt, Spitze transformiert in männliches Blüten-Primordium, Wurzel, unreife männliche Blüte
MYBPH3_M00218	196	unterschiedlich reguliert: Nematoden, 6-Benzyl-Adenin, Chitin, CO2_hoch, Norflurazon, Syringolin, Hormon IAA, Seneszenz, Wärme	keine Gemeinsam-keiten	keine Gemeinsam-keiten	keine Gemeinsam-keiten
P	5	unterschiedlich reguliert: Ozon, wenig-N-Glucose, wenig Nitrat, Osmose	Expression: ausgereifte Schoten	keine Gemeinsam-keiten	involviert in Flavonoid-Metabolismus
PIF3(2)	227	upreguliert: weiter entferntes Rot; unterschiedlich reguliert: Nematoden,	Expression: **ausgereifte Schoten**	keine Expression: **Callus, Keimwurzel, Wurzeln**	Expression: **ausgereifte Schoten**; keine Expression: **Callus, Keimwurzel, Wurzeln**; keine

		Norflurazon, wenig-N-Glucose, Seneszenz, Osmose			Übereinstimmung: weiter entferntes Rot
RAV1_ M00344	4	unterschiedlich reguliert: Cycloheximid, Wärme (31 von 32)	Expression: ausgereifte Schoten	Expression: Stängel	keine Gemeinsam-keiten
TATABP	0	unterschiedlich reguliert: Chemikalische Komponenten	keine Gemeinsam-keiten	keine Gemeinsam-keiten	Vergleich nicht notwendig
TEIL_ M00502	5	keine Gemeinsam-keiten	keine Gemeinsam-keiten	keine Gemeinsam-keiten	Vergleich nicht notwendig

Tabelle 53: Exon3: Auswahl der niedrigsten 10 P-Werte und Auflistung der Expression der entsprechenden putativen Zielgene. Für die Expressionsdaten der putativen Zielgene wurde Genevestigator eingesetzt. Expressionsdaten bezüglich des Transkriptionsfaktors wurden mittels TRANSFAC© sowie Genevestigator, wenn der Transkriptionsfaktor aus *Arabidopsis thaliana* stammt, festgestellt. Übereinstimmungen der Expressionsdaten der putativen Zielgene und des Transkriptionsfaktors sind fett gedruckt.

TF	**Distanz**	**Stimulus**	**Development**	**Anatomy**	**TF-Expression**
ABI4(2)	32	unterschiedlich reguliert: 6-Benzyl-Adenin, Syringolin, wenig-N-Glucose, Kälte	keine Gemeinsam-keiten	keine Gemeinsam-keiten	Vergleich nicht notwendig
AG_ M00151	4	unterschiedlich reguliert: Isoxaben, Kälte	keine Gemeinsam-keiten	keine Gemeinsam-keiten	Vergleich nicht notwendig
AGP1	5	unterschiedlich reguliert: Isoxaben, TIBA, Wärme	Expression: ausgereifte Schoten	Expression: Stängel	pos: Blätter, suspensionskultivierte Zellen, Blüte, Wurzel

AtMYB15	3	unterschiedlich reguliert: Cycloheximid, Wärme	Expression: ausgereifte Schoten	Expression: Stängel	keine Gemeinsam-keiten
AtMYB84	3	unterschiedlich reguliert: Cycloheximid, Wärme	Expression: **ausgereifte Schoten**	keine Gemeinsam-keiten	Expression: **ausgereifte Schoten**
bZIP910(2)	43	downreguliert: Hormon MJ; unterschiedlich reguliert: 6-Benzyl-Adenin, Cycloheximid, Salicylsäure, Licht, Hypoxie	Expression: ausgereifte Schoten	Expression: Stängel; keine Expression: Fruchtknoten, Stigma, Wurzel-komponenten	Blüten, Wurzeln, Blätter
P	5	unterschiedlich reguliert: 2,4 Dichlorophenyoxalat	Expression: ausgereifte Schoten	Expression: Blüte	involviert in Flavonoid-Metabolismus
RAV1_ M00344	4	keine Gemeinsamkeiten	keine Gemeinsam-keiten	keine Gemeinsam-keiten	Vergleich nicht notwendig
TATABP	0	keine Gemeinsamkeiten	keine Gemeinsam-keiten	keine Gemeinsam-keiten	Vergleich nicht notwendig
TEIL_ M00502	5	keine Gemeinsamkeiten	keine Gemeinsam-keiten	keine Gemeinsam-keiten	Vergleich nicht notwendig

Tabelle 54: Exon4: Auswahl der niedrigsten 10 P-Werte und Auflistung der Expression der entsprechenden putativen Zielgene. Für die Expressionsdaten der putativen Zielgene wurde Genevestigator eingesetzt. Expressionsdaten bezüglich des Transkriptionsfaktors wurden mittels TRANSFAC© sowie Genevestigator, wenn der Transkriptionsfaktor aus *Arabidopsis thaliana* stammt, festgestellt. Übereinstimmungen der Expressionsdaten der putativen Zielgene und des Transkriptionsfaktors sind fett gedruckt.

TF	**Distanz**	**Stimulus**	**Development**	**Anatomy**	**TF-Expression**

AG_ M00151	4	unterschiedlich reguliert: Cycloheximid, Seneszenz	Expression: Keimling, junge Rosette, sich entwickelnde Rosetten, bolting, junge Blüte, sich entwickelnde Blüten, **Blüten und Schoten**, ausgereifte Schoten	Expression: Stängel, Keimling (und components), Rosette (und Komponenten)	Expression: **Blüten und Schoten**; keine Gemeinsamkeiten: Stängel, Keimling, Rosette
AGP1	5	downreguliert: Seneszenz; unterschiedlich reguliert: TIBA, Wärme	Expression: ausgereifte Schoten	Expression: Stängel	pos: Blätter, suspensionskultivierte Zellen, Blüte, Wurzel
AtMYB15	3	unterschiedlich reguliert: Wärme	Expression: ausgereifte Schoten	Expression: Stängel	keine Gemeinsam-keiten
AtMYB15	71	unterschiedlich reguliert: Licht, wenig-N-Glucose, wenig Nitrat, Anoxie, Wärme	Expression: ausgereifte Schoten	Expression: Hypocotyl	keine Gemeinsam-keiten
AtMYB77	70	unterschiedlich reguliert: *A. tumefaciens*, Syringolin	keine Gemeinsam-keiten	keine Gemeinsam-keiten	Vergleich nicht notwendig
GAMYB	183	1 Arrayelement für alle: durch Chemikalien,	1 Array-element überall keine Expression	1 Array-element Expression: Staubgefäße,	pos: Aleuronschicht; Staubgefäße, Ähren; Anthere, Epidermis, sich entwickelnde

		Stress, Licht, Nährstoffe, Hormone, biotischen Stress upreguliert bzw. downreguliert		Pollen	Samen, Endosperm, vaskuläres Gewebes des Scutellums, Trieb des Weiterbildungsgewebes und Keimwurzel; neg: Spelze/Lemma, Granne, Fruchtknoten, Spelze, Blatt, Wurzeln; junge Blätter und Wurzeln
HVH21	1	unterschiedlich reguliert: Seneszenz	Expression: ausgereifte Schoten	Expression: Stängel	pos: Lemma, Spelze, vaskuläres Bündel , basale Region inclusive des SAM
RAV1_ M00344	4	unterschiedlich reguliert: Isoxaben	Expression: **junge Rosette**, sich entwickelnde Rosetten	keine Gemeinsamkeiten	Expression: **junge Rosette**
TATABP	0	unterschiedlich reguliert: Ibuprofen, Syringolin	Expression: ausgereifte Schoten	keine Gemeinsamkeiten	ubiquitär
TEIL_ M00502	5	unterschiedlich reguliert: Wärme	keine Gemeinsamkeiten	keine Gemeinsamkeiten	Vergleich nicht notwendig

Tabelle 55: 3'UTR1: Auswahl der niedrigsten 10 P-Werte und Auflistung der Expression der entsprechenden putativen Zielgene. Für die Expressionsdaten der putativen Zielgene wurde Genevestigator eingesetzt. Expressionsdaten bezüglich des Transkriptionsfaktors wurden mittels TRANSFAC© sowie Genevestigator, wenn der Transkriptionsfaktor aus *Arabidopsis thaliana* stammt, festgestellt. Übereinstimmungen der Expressionsdaten der putativen Zielgene und des Transkriptionsfaktors sind fett gedruckt.

TF	**Distanz**	**Stimulus**	**Development**	**Anatomy**	**TF-Expression**

ATHB1	99	unterschiedlich reguliert: Seneszenz, Salz (6 von 7); upreguliert: Wärme (6 von 7)	Expression: ausgereifte Schoten	Expression: Stängel	keine Gemeinsam-keiten
ATHB1	89	unterschiedlich reguliert: *A. tumefaciens, B. cenerea*, Nematoden, einigen Chemikalien, z.B. Chemikalie Ibuprophen; downreguliert: Seneszenz (4 von 5)	Expression: ausgereifte Schoten	Expression: Stängel	keine Gemeinsam-keiten
ATHB1	2	downreguliert: Kältestress; unterschiedlich reguliert: Hormon BL/H_3BO_3 (stark!), Chemikalie Isoxaben, Glucose, wenig-N-Glucose, Seneszenz	keine Gemeinsam-keiten	keine Gemeinsam-keiten	keine Gemeinsam-keiten
ATHB5	6	unterschiedlich reguliert: Hormon BL/H_3BO_3 (stark!), wenig-N-Glucose (6 von 7 downreguliert), Seneszenz	keine Gemeinsam-keiten	Expression: **Endodermis**	Expression: **Endodermis**
ATHB5	208	downreguliert: Chemikalie **Syringolin**; unterschiedlich reguliert: Licht, wenig-N-Glucose, Wärme	keine Gemeinsam-keiten	keine Gemeinsam-keiten	stark downreguliert: Chemikalie **Syringolin**
ATHB9_ M00417	10	keine Gemeinsam-keiten	keine Gemeinsam-	keine Gemeinsam-	keine Gemeinsam-

			keiten	keiten	keiten
CDC5_ M00361	2	unterschiedlich reguliert: *A. tumefaciens*, Chemikalie Cycloheximid, Isoxaben, Norflurazon, Lichtintensität, UV-Filter, Stress Anoxie, Kälte, Wärme	Expression: **ausgereifte Schoten**	Expression: **Stängel**	Expression: **ausgereifte Schoten, Stängel**
HAHB4	5	downreguliert: Syringolin; unterschiedlich reguliert: *B. cinerea, P. infestans*, Chemikalie 2,4-Dichlorophenoxylat, AgNO3, AVG, CO3 hoch, Cycloheximid, Daminozid, Isoxaben, Norflurazon, NPA, Ozon, Nährstoffe, wenig-N-Glucose, Saccharose, Stress Anoxie, Wärme, Hypoxie, Osmose	keine Gemeinsam-keiten	keine Gemeinsam-keiten	unbekannt
MYBPH3_ M00218	8	unterschiedlich reguliert: wenig Nitrat, Seneszenz, Anoxie	keine Gemeinsam-keiten	keine Gemeinsam-keiten	Vergleich nicht notwendig
MYBPH3_ M00219	8	unterschiedlich reguliert: $AgNO_3$, Cycloheximid, Hormon BL/H_3BO_3, Wärme	keine Gemeinsam-keiten	keine Gemeinsam-keiten	Vergleich nicht notwendig

Tabelle 56: Auswahl der niedrigsten 10 P-Werte und Auflistung der Expression der entsprechenden putativen Zielgene bezüglich des intergenischen Bereichs, unter der Voraussetzung, dass die zu den putativen Transkriptionsfaktor-Bindungsstellen zugeordneten Gene einen annotierten Transkriptionsstartpunkt aufweisen. Unberücksichtigt blieb hierbei der Transkriptionsfaktor TBP, da dieser erwartungsgemäß eine Vielzahl von Bindungsstellen im intergenischen Bereich aufweist. Für die Expressionsdaten der putativen Zielgene wurde Genevestigator eingesetzt. Expressionsdaten bezüglich des Transkriptionsfaktors wurden mittels TRANSFAC© sowie Genevestigator, wenn der Transkriptionsfaktor aus *Arabidopsis thaliana* stammt, festgestellt. Übereinstimmungen der Expressionsdaten der putativen Zielgene und des Transkriptionsfaktors sind fett gedruckt.

TF	**Distanz**	**Stimulus**	**Development**	**Anatomy**	**TF-Expression**
ABF1	-78	unterschiedlich reguliert: Chemikalie Dexamethason, Hormon ABA	Expression: ausgereifte Schoten	keine Gemeinsam-keiten	keine Gemeinsam-keiten
ABF1	-88	unterschiedlich reguliert: Hormon ABA, Zeatin	keine Gemeinsam-keiten	keine Gemeinsam-keiten	kein Vergleich notwendig
O2	-92	keine Gemeinsam-keiten	Expression: ausgereifte Schoten	keine Gemeinsam-keiten	pos: Endosperm
O2	-66	unterschiedlich reguliert: Wärme	keine Gemeinsam-keiten	keine Gemeinsam-keiten	keine Gemeinsam-keiten
O2	-88	keine Gemeinsam-keiten	Expression: ausgereifte Schoten, Blüten und Schoten	Expression: Stängel	pos: Endosperm
O2	-79	unterschiedlich reguliert: Cycloheximid, Seneszenz	Expression: **ausgekeimte Samen**	keine Gemeinsam-keiten	pos: **Endosperm**
ABF1	-85	unterschiedlich reguliert: *B. cinerea*, Chemikalie: Syringolin, Hormon ABA	keine Gemeinsam-keiten	keine Gemeinsam-keiten	kein Vergleich notwendig

O2	-105	unterschiedlich: 6-Benzyl-Adenin, Dexamethason, NAA, Zeatin, Glucose, Wärme, osmotischer Stress, Salz	keine Gemeinsam-keiten	keine Gemeinsam-keiten	kein Vergleich notwendig
O2	-72	unterschiedlich reguliert: ABA, Wärme, Osmose, Salz	keine Gemeinsam-keiten	keine Gemeinsam-keiten	kein Vergleich notwendig
ABF1	-66	unterschiedlich: Chemikalie: Cycloheximid, Dexamethason, Hormon ABA, Glucose, Nitrat-Saccharose	keine Gemeinsam-keiten	keine Gemeinsam-keiten	kein Vergleich notwendig

Tabelle 57: Auswahl der niedrigsten 10 P-Werte und Auflistung der Expression der entsprechenden putativen Zielgene bezüglich des intergenischen Bereichs, unter der Voraussetzung, dass die zu den putativen Transkriptionsfaktor-Bindungsstellen zugeordneten Gene lediglich einen annotierten Translationsstartpunkt und keinen Transkriptionsstartpunkt aufweisen. Für die Expressionsdaten der putativen Zielgene wurde Genevestigator eingesetzt. Expressionsdaten bezüglich des Transkriptionsfaktors wurden mittels TRANSFAC© sowie Genevestigator, wenn der Transkriptionsfaktor aus *Arabidopsis thaliana* stammt, festgestellt. Übereinstimmungen der Expressionsdaten der putativen Zielgene und des Transkriptionsfaktors sind fett gedruckt.

TF	**Distanz**	**Stimulus**	**Development**	**Anatomy**	**TF-Expression**
AGL15	-8	upreguliert: Chemikalie PCIB; unterschiedlich reguliert: *A. tumefaciens,* $AgNO_3$, Wärme	Expression: Keimling, junge Blätter, sich entwickelnde Rosetten	keine Gemeinsam-keiten	downregulation bei PCIB
ATHB5	-4	unterschiedlich reguliert: 2,3-Dichlorophenoxalat, wenig Nitrat,	keine Gemeinsam-keiten	keine Gemeinsam-keiten	kein Vergleich notwendig

		Wärme			
AtMYB15	-56	downreguliert: Chemikalie **AVG**; unterschiedlich reguliert: *A. tumefaciens*, Mycorrhiza, Nematoden, Saccharose, Seneszenz, Anoxie, Wärme, Hypoxie	Expression: ausgereifte Schoten	Expression: Zellsuspension, Stängel	downreguliert bei **AVG**, ansonsten keine Gemeinsam-keiten
GAMYB	-502	1 Arrayelement: downreguliert: vorwiegend bei Chemikalien, upreguliert: vorwiegend bei Licht, unterschiedlich vorwiegend bei Stress, Hormon	1 Array-element, überall Expression	1 Array-element Expression: Stängel, seitliche Wurzeln, Leitbündel, Staubgefäße, Kronenblatt	pos: Aleuronschicht; Staubgefäße, Ähren; Anthere, Epidermis, sich entwickelnde Samen, stärkehaltiges Endosperm, Kernauslagerung, vaskuläres Gewebes des Scutellums, Trieb des Bildungsgewebes und Keimwurzel; neg: Spelze/Lemma, Granne, Fruchtknoten, Spelze, Blatt, Wurzeln; junge Blätter und Wurzeln
HAHB4	-232	unterschiedlich reguliert:	keine Gemeinsam-	Expression: Stängel	unbekannt

		Nematoden, Cycloheximid, TIBA, Hormon BL/H_3BO_3, Seneszenz, Anoxie, Kälte, Wärme, Osmose	keiten		
MYBPH3_ M00218	-85	downreguliert: Norflurazon, upreguliert: Anoxie; grün: Trockenheit; unterschiedlich reguliert: *A. tumefaciens*, Nematoden Paclobutrazol, PCIB, Chitin, Salicylsäure, Hormon Zeatin, Wärme	keine Gemeinsam-keiten	Expression: Zellsuspension	pos: Blüte; neg: Blätter und Stängel, Kelchblätter, Wurzeln
MYBPH3_ M00218	-204	nicht repräsentativ	nicht repräsentativ	nicht repräsentativ	kein Vergleich notwendig
RAV1_ M00343	-13	nicht repräsentativ	nicht repräsentativ	nicht repräsentativ	kein Vergleich notwendig
RAV1_ M00343	-9	nicht repräsentativ	nicht repräsentativ	nicht repräsentativ	kein Vergleich notwendig
RAV1_ M00343	-687	downregulation: Chemikalie **Propiconazol**, **Uniconazol**; unterschiedlich: ABA, Zeatin, Saccharose; upregulation: **Wärme**	keine Gemeinsam-keiten	keine Gemeinsam-keiten	downreguliert: **Propiconazole**, **Uniconazole**; upreguliert: **Wärme**

Danksagung

Ich bedanke mich bei allen Personen, die mich bei der Durchführung meiner Doktorarbeit unterstützt haben.

Bei Prof. Dr. Reinhard Hehl bedanke ich mich für die Überlassung des Themas, für die Diskussionsbereitschaft bezüglich wissenschaftlicher Fragestellungen und die Förderung, wissenschaftliche Ergebnisse präsentieren zu können. Des Weiteren bedanke ich mich dafür, dass er mir die Teilnahme an einem internationalen Kongress ermöglicht hat.

Mein Dank gilt auch Prof. Dr. Dieter Jahn für die freundliche Übernahme des Koreferats.

Ich bedanke mich bei Dr. Lorenz Bülow für die Diskussionen und für die Entwicklung der TFBSitesTools, die den Anlass für das Thema meiner Doktorarbeit gaben.

Bei allen Mitarbeitern des Instituts für Genetik möchte ich mich für die angenehme Arbeitsatmosphäre bedanken. Insbesondere bei Dr. Heiko Oltmanns und Dr. Nils Ole Steffens für ihre Freundschaft und ihre Fähigkeit, Menschen zu begleiten.

Mein besonderer Dank gilt Herrn Dr. José L. Lozán für das Testen von Zufallspositionen auf Normalverteilung.

Herzlich bedanken möchte ich mich bei den Mitarbeitern des Instuts für Mathematische Stochastik an der Technischen Universität Braunschweig für die hilfreichen Anregungen bezüglich der Verteilungen von Transkriptionsfaktor-Bindungsstellen.

Besonders bedanke ich mich bei meiner Familie.

Diese Arbeit entstand im Rahmen des Intergenomics Netzwerks, gefördert vom Bundesministerium für Bildung und Forschung, BMBF, Förderkennzeichen 031U110C / 031U210C

Lebenslauf

Name:	Claudia Galuschka
Geburtsdatum:	31.08.1971
Geburtsort:	Wickede-Wimbern
Familienstand:	ledig
Staatsangehörigkeit:	deutsch

Schulausbildung und Studium

1978 - 1982	Grundschule in Letmathe
1982 - 1988	Realschule in Letmathe Abschluss: Fachoberschulreife
1988 - 1989	Vorpraktikum im ev. Kindergarten Oestrich
1989 - 1991	Fachschule für Sozialpädagogik Abschluss: Staatlich geprüfte Erzieherin
1991 - 1992	Anerkennungsjahr im kath. Kindergarten Oestrich Abschluss: Staatlich anerkannte Erzieherin
1993 - 1996	Abendgymnasium Dortmund Abschluss: Allgemeine Hochschulreife
1996 - 2003	Studium der Biotechnologie an der Technischen Universität Braunschweig Abschluss: Diplom-Biotechnologin

Berufsausübung

1992 - 1996	Erzieherin im kath. Kindergarten Oestrich
2003 - 2007	Wissenschaftliche Mitarbeiterin (Doktorandin) an der Technischen Universität Braunschweig, Institut für Genetik
2008-	Projektleiterin Datenbanken beim Deutschen Krebsforschungsinstitut (DKFZ) in Heidelberg